MIT Schwarzman College of Computing Task Force

Working Group on Computing Infrastructure

Final Report

Nicholas Roy and Benoit Forget, Co-Chairs

August 5, 2019

OVERVIEW

The design, development, and deployment of resources to support research and education across campus requires a deep understanding of how computing is used within the disciplines. It also requires intellectual leadership from both the computing research communities and the individual disciplines integrated with an institutional commitment to a robust operational infrastructure. Currently, MIT has no sustained institutional commitment to a robust infrastructure and, with a few exceptions, the intellectual leadership has not been deeply integrated with the operational capabilities on campus.

Computing infrastructure is a sustained institution-wide combination of governance, people, funding processes, hardware, and software. It addresses and serves education and research as well as access needs across all disciplines encompassed by MIT's academic endeavors. Our working group found a strong view across campus that the full success of the Schwarzman College of Computing (SCoC) will depend on creating a solid and robustly sustained computing infrastructure that is MIT-wide with strong engagement of academic and operational perspectives. In weighing up pros and cons, our working group found a campus-wide preference for an overarching organizational model of computing infrastructure that transcends a college or school and most logically falls under senior leadership.

We observed enthusiasm and broad support for re-examining how research computing is deployed across campus. Robust infrastructure is crucial not just for the SCoC but for all of MIT. We are excited about the opportunity to change how we compute across disciplines

and to change the very culture of equitable access to computing. Our goal is to begin developing a common framework and language for thinking about computing across campus that ensures equitable access to resources.

Our working group was charged with examining how to ensure that departments have the information and resources they require to meet their computational needs, including methods of accessing and storing data. Among the questions that guided our exploration:

- What is the best way to assess the need for computing infrastructure?
- How should computing resources be distributed within the SCoC and throughout MIT?
- How should these resources be maintained and renewed?

Our working group surveyed representatives from most DLCs and a few student groups. Members from our working group met with these stakeholder groups to gather information about their computing status, their current and future needs, as well as the major pain points they experience with computing resources. In addition to these meetings, we also leveraged prior work on campus related to computational thinking, data for healthcare applications, and a survey on research computing. We also monitored the idea bank from the SCoC website. We organized weekly meetings with invited guests to discuss the computing landscape at MIT and at peer institutions, hardware and software needs, and data licensing and access issues.

The major gaps we found in the course of our research include:

- A current imbalance between centralized and decentralized models that is no longer working and is creating substantial inefficiencies and inequities.
- A lack of support for data management that is broadly considered a major limitation inherent in research on campus.
- The unsustainability of the current operating model.
- No easily accessible education-focused computing resources.
- A lack of focus on both long-term corporate memory and a plan for renewal has led to substantial limitations and inefficiencies, currently and in the past.

The major needs we identified in the course of our research include:

- A reexamination of how research computing is deployed across campus—a need that is broadly and enthusiastically supported.
    - No single model will work for the diverse needs of the campus, so MIT must remain flexible.
    - It is essential to reexamine the balance between decentralized, centralized, and cloud computing.
    - A focus on accessibility and equity for research and education is important.
- Support for infrastructure that includes appropriate levels of professional staffing, community building, student support, and training support, including:
    - Professional service model for research and education support.
    - Student-based support (e.g., Athena).
    - IAP seminars, workshops, etc.
- A framework for bringing data to campus and managing it equitably, securely, and responsibly.
    - Oversight organization is needed to facilitate data acquisition and management.
- Funding and budgeting mechanisms for a sustainable and renewable model of computing across campus.
    - It will be important to find the right balance between PI funding, department funding, Institute funding, and hardware or cloud credit donations.
    - Computing infrastructure will require a commitment from both the SCoC and MIT at-large to meet the needs of the campus.
- Strong interest exists across campus for the advancement of environmentally sustainable models of computing.
    - This interest exemplified by the Massachusetts Green High Performance Computing Center (MGHPCC).
    - There also exists interest in upholding the cloud providers to our high standards.
- Equally strong interest exists for finding ways to improve equitable access to resources across the campus.
    - Facilitate access for education.
    - Facilitate access to DLCs not typically associated with computing.

See the Institutional Analysis section beginning on p. 29 to review our assessment of the current needs, resources, and challenges of MIT's computing infrastructure.

KEY IDEAS

EXISTING MODELS

We examined the current state of computing on campus, gathering information from a recent survey by the Committee on Research Computing, from surveys performed by the working group with DLCs, and from information gathered via white papers. We have presented our findings in terms of:

- organizational models
- resource deployments
- data management and data licensing
- funding/resource models

Organizational Models

At present, research computing at MIT has three overlapping styles of activity:

- individual principal investigator (PI-based approach)
- departmental collaborations
- institutional efforts that are open to everyone

In each case, we cite examples of either physical hardware or commercial cloud-based systems that are being used to meet needs, as well as instances that mix physical resources and cloud resources.

Individual PI-Based Models

Individual PI-based models involve computing resources and associated support activities operated by a PI for the sole use of an individual research group. These can range from single servers to multi-rack clusters (or their virtual equivalent). The resources are typically funded either from discretionary funds or from sponsored research funds. For PIs with adequate budgets to sustain the hardware aspects, the associated support costs, and the necessary time commitments, this can be an attractive approach.

In theory, the individual model allows the most flexibility and autonomy, but economies of scale can be difficult to realize. For some research groups, this is an important factor—especially for groups with specialized hardware or hardware-access needs. For some PIs, however, maintaining enough funding to sustain this approach long term and as needs grow or change can become a burden. In the case of physical resources, costs for space, power, and cooling fall to departmental and institutional budgets. Administration of PI-specific computing resources is usually carried out by members of the PI's group. A modest number of these individual-model efforts operate exclusively with remote resources (either government or industry collaborator resources or commercial cloud resources).

All the commercial cloud providers have been keen to provide some credits to individual PI groups. The amount of the credits varies considerably depending on cloud provider interest in the type or nature of the research. This has enabled some significant computational research projects and also has supported many classes. What resources are available, however, and to whom and when those resources are available are all somewhat dependent upon external factors.

Many individual groups also make use of off-campus facilities such as those supported by the federal government or by industry partners. A number of groups make use of a range of resources across MIT, working with commercial cloud providers and at national computing centers.

Often, PI resources are leveraged to support educational efforts offered by the PI, but these resources are not always accessible to others teaching the same course.

Departmental Collaborations

Some DLCs maintain resources that are shared across multiple PIs. These resources are typically larger in scale than resources operated by individual Pis, and access to them is usually limited to a specific group of PIs. This approach facilitates some sharing of costs and associated economies and also facilitates accessibility for education.

The departmental approach also allows the customization of resources to meet the specialized needs of a particular style of research. For example, Brain and Cognitive Sciences operates a very dense, highly cost-optimized, consumer-grade GPU cluster that is tuned to the needs of a specific set of neuroscience applications. CSAIL, in turn, operates an OpenStack service that is geared toward the experimental needs of that community.

The scale of typical departmental computing resources requires support staff to administer and manage. Some groups—CSAIL, for example—support dedicated staff teams to maintain their services. Others, such as Brain and Cognitive Sciences, rely upon a combination of dedicated in-house staff and outside consulting. This approach is favored by larger DLCs that can sustain the associated costs.

The hardware and system administration costs associated with these systems are largely funded through the pooling of departmental and PI group charges. Some departments have a tiered funding model, allowing PIs to buy in at different levels depending on their usage. Others allow PIs to purchase computing and storage that is available for their use when needed but that can otherwise be shared across the group.

Shared departmental resources tend to be larger than individual PI resources. All these resources are physical systems and are located at off-campus data centers. In terms of off-campus resources, the costs associated with space, power and cooling, and providing physical access to remote hands are supported by central administration budgets. CSAIL and the Laboratory for Nuclear Science both operate sizable resources as departmental facilities. Some departmental resources are maintained in campus spaces. Biology, for example, operates an on-campus facility that supports sequencing and other laboratory needs. A few departmental facilities are maintained partly because of tight data systems and physical lab instrument coupling needs or licensing or networking restrictions that preclude participation in Institute-wide shared pools. Successful departmental models usually have some low-overhead governance structure to gather feedback and ensure that the system is meeting needs (e.g., user meetings).

At the departmental/single organization level, commercial cloud providers such as AWS, Azure, Google, and IBM have provided some credits to strategic activities in teaching and research and around modest scale data storage for meeting library needs. Such arrangements have benefited individual DLCs.

Institutional Efforts

The largest physical systems at MIT are operated under an institutional model. Resources operated under this model are available to anyone at MIT regardless of their DLC associations. The operating costs of these systems are supported centrally. The hardware in these resources is still nearly all funded by individual PI discretionary and sponsored research funds (with the exception of Lincoln Laboratory). Institutional-model supported systems include collaborative initiatives with Lincoln Laboratory. Groups using these systems daily span researchers from nearly all DLCs and cross many disciplines and usage models.

System administration and support costs of the institutional model are primarily covered by funds from the offices of the Vice President for Research (VPR) and Provost. The resources provide the largest economies of scale and represent the most extensive integrated systems available on campus. The institutional-model systems support research and classroom activities for a wide range of groups. In 2018, the systems collectively supported approximately 500 million core hours of computation and housed approximately 10 PiB of data. The scale of these resources tends to make their hardware somewhat less flexible than departmental-level resources.

Through a mix of sponsor support and VPR/Provost funding, current institutional-model systems provide a set of base services to MIT researchers for no direct charge. PIs also purchase additional computing and storage resources to go beyond base needs and can contract with support staff for additional intensive assistance beyond basic ticketing support. Some DLCs (for example MIT Sloan and the MIT Plasma Science and Fusion Center) have local support staff teams to augment the central service. For DLCs that can justify it, this can be a particularly effective way to balance economies of scale with department-specific custom needs.

Currently, all the institutional-model production resources are based on physical hardware rather than commercial cloud platforms, primarily for cost reasons. Commercial cloud counterparts could offer increased flexibility, scalability, and quality of service for much of the supported workload. Several proof-of-concept exercises have been undertaken to demonstrate this. Present market prices for commercial cloud approaches mean that, absent a significant donation, a cloud-based approach would increase institutional and PI total costs significantly. A number of departmental efforts are experimenting with sharing some workload with Institute-wide resources, which can further optimize costs. Institutional-model activities at MIT also include a modest amount of training and support activities (currently limited by available resources). These training activities include assistance with using local resources and with getting access to national facilities.

A recent experiment with cloud credits from Google and IBM (managed by the Bridge under the Quest for Intelligence) is using an open-to-all institutional model. The Google cloud credits have proven quite popular, and proof-of-concept studies have shown how it can support interactive workloads that refer back to data on current-generation, institutional-model production systems. The Google credits have allowed several groups to access sizable GPU resources for numerical modeling at levels not possible with campus resources. Unfortunately, the current Google cloud credit allocation expires August 31, 2019, and the follow-on resources are still being negotiated.

The IBM cloud credits controlled by the Bridge also are proving useful and are likely to run for a more extended period. As with other commercial clouds, it supports the evolving Kubernetes standard. This allows it to form the hardware basis of scalable, interactive, open-source notebook services that are increasingly becoming an important resource for teaching and research tool development as well as en masse classroom deployment (subject to the availability of content and suitably trained teaching staff). IBM Cloud also is used heavily to support the MIT–IBM contribution to the MIT Quest Core work.

Institutional initiatives planned in 2019 and 2020 will explore hybrid cloud models with IBM and Redhat. The hybrid cloud approach can extend successful cloud experiments to

create more integration across on-premise and cloud resources. The planned hybrid cloud initiatives will be enabled by a sizable IBM hardware gift announced at the SCoC launch event. In addition to adding valuable resources to address seemingly insatiable campus AI research needs, the multi-million dollar IBM gift will support practical work in developing hybrid cloud models in which workloads and data move more easily between MIT facilities and elastic cloud environments such as the IBM cloud.

Institutional-model activities at MIT are guided by a governance group that has faculty representation from all schools and that advises the VPR office. The governance group participates in the Institute Information Technology Governance Council, which helps prioritize budgeting and cost optimization decisions.

Organizational Models Pros and Cons

All three organizational models have merit, and we anticipate that overall future infrastructure needs might continue to be met through some combination of these. All models can exist as either physical hardware-based environments, as commercial cloud-based environments, or as some combination.

Individual PI-Based Model Pros and Cons

The individual model provides the most flexibility to an individual PI. For this reason, this model has real value. Taken to the limit with physical systems, however, it can lead to a proliferation of resources that are optimal for a particular goal but vastly suboptimal in terms of overall impact, space, power, cooling, cybersecurity, support staff, duplicative effort, campus planning, and facilities infrastructure management. Over time, relying on this approach exclusively would become prohibitively demanding on building facilities resources.

Commercial cloud resources could, in theory, mitigate many of these factors and provide a tremendously flexible and capable platform for meeting individual needs. MIT has approximately 2,800 accounts on AWS, and many students take advantage of its free-tier computing. This could make individual-model approaches that leverage commercial cloud resources potentially irresistible. Unfortunately, at present cloud

prices, scenarios that involve sustained computation or large data storage do not make economic sense for many PIs. A 2019 survey of campus PIs revealed that 50 percent of PIs viewed cost as a reason for not using commercial cloud services.

Another challenge of an exclusively individual model is that it requires expertise within a PI's lab to know how to acquire, set up, secure, use, maintain, and protect computing resources. As data sets with limiters grow (HIPPA, IRB, proprietary), the expertise required to manage computing resources in a legally compliant manner becomes significant. It is less clear how an exclusively individual model would result in cost-effective provisioning of equitable access and support for both research and education to all of MIT. Ignoring this problem could impact overall institutional productivity, competitiveness, and ranking as well as shrink the potential transformative impact of the SCoC.

Departmental Model Pros and Cons

The departmental model, when well supported, allows groups to pool funds and build or configure systems to meet very particular group needs. It can be a very effective model, and multiple successful departmental-model systems are now operating at MIT. The departmental model provides these groups with a degree of autonomy and control. That said, an equal number of departmental-model systems at MIT have proved unsustainable. A common challenge with these systems is the need for stable funding for support staff to maintain an adequate quality of service. Another challenge can be sustaining long-term commitment across a collaborative group. This model works well for groups with well-funded and stable base-research revenues. In less robustly funded groups, the departmental model can be hard to sustain when group leadership changes and group interests shift.

Institutional Model Pros and Cons

The institutional model can, in principle, be very effective at ensuring that everyone has access to some level of computational resources and that access to the greatest number of resources is possible. To fully realize its potential, this model requires robust funding to allow for stable multi-year planning. The institutional model can be implemented with physical hardware or through commercial cloud resources or through a mix. At

present, commercial cloud resources at market prices are the most expensive approach. Large donations or long-term cost agreements for cloud resources can potentially make this model very low cost in terms of facilities impact and meeting operational demands. The MIT IBM partnership that includes combined donations of sizable hardware (as announced at the SCoC launch event) and cloud credits (through the MIT–IBM Quest) is a compelling example of potential partnerships.

The institutional model maximizes institutional economies of scale and allows infrastructure management to balance individual PI and DLC priorities with overall institutional priorities. This can create disconnects between research growth (i.e., increased needs for space, power and cooling) and the Institute's capacity to meet demands. The institutional model tends to shift costs toward central administration and department funds and away from PI direct funds. This often reduces the total absolute cost, thus making it the lowest cost option in absolute terms. However, it also shifts where those costs fall in ways that can create concerns at the institutional budgeting level. Some of this can be offset with charges for services beyond some baseline amount. Almost inevitably, however, this model looks more expensive to central administration activities than other models, despite having the lowest absolute costs.

Overall, each of these models has a valuable role in a vibrant research computing ecosystem. Each has the potential to provide an efficient degree of cost-sharing and collaboration. In the fast changing and volatile research computing world, a diversity of approaches can be an asset if effectively channeled.

Resource Deployments

Various groups have retained a number of departmental server rooms that must support server clusters on campus, either because of access needs or, in some cases, because of a lack of space off campus.

To support diverse, large-scale needs beyond campus, MIT currently owns, partially owns, or leases several sizable data center facilities along with a significant wide-area networking infrastructure. Together, these provide key pieces for enabling a cost-effective environment

for both physical and cloud-based services. The wide-area network capabilities available to MIT consist of 72 dark fiber pairs that have the potential to provide unprecedented bandwidth to MIT data centers as well as to cloud provider and commercial and academic Internet access points. The MIT off-campus data center facilities include an enterprise class ~400KW facility in downtown Boston, an ~1MW facility at the Bates Laboratory in Middleton, and a potentially 35MW shared facility in Holyoke.

Taken together, these facilities provide an excellent foundation for growing physical hardware capacity (subject to appropriate facility upgrades) and for developing high-speed pathways from research offices and labs to commercial cloud services as well as to remote facilities. MIT researchers use these facilities to house physical resources and to access cloud resources and national research facilities 24/7, 365 days a year.

In a survey of computing needs, approximately 75 percent of respondents reported using physical resources either on campus or at off-campus data centers. Roughly 30 percent of respondents reported using commercial cloud resources, and about 30 percent reported using various national facilities or other non-commercial, non-MIT resources for large-scale computing.

At the software level, we see a growing convergence between interactive environments and large-scale computing environments—as well as between physical hardware environments and cloud environments. Many resources that researchers use today support browser-based interaction through tools such as Jupyterhub and Rstudio, graphical platforms such as Matlab, and batch-scheduled, high-throughput, and high-performance computing all in one system. All these services are available over networks in the cloud or on local clusters.

Resource Deployments Pros and Cons

Physical facility deployments, although requiring more infrastructure, are significantly less costly overall than commercial cloud resources for intensive use. Commercial cloud resources, on the other hand, are significantly more flexible and adaptable for many problems and very cost effective for less intensive use. However, both cloud-only and physical-facility-only models have significant drawbacks. Hybrid models that fuse cloud and physical resources to manage costs as seamlessly as possible appear attractive to many MIT

groups at the moment, and hybrid models have been developing rapidly in the last few years. Key elements such as allowing common identities across systems and data sharing now have real-world working examples that make this possible. Increasingly, it is possible to have software stacks and environments that are indistinguishable to end users, regardless of their locations.

Data Management and Data Licensing

Research teams at MIT are collectively acquiring, processing, analyzing, and storing highly sensitive data at extreme scale on a daily basis. Yet, MIT has a troubling lack of Institute-wide infrastructure, standards, and policies for research data management and equitable, centrally-accessible infrastructure for data storage and sharing. The Institute's highly distributed infrastructure, standards, and policies for research data-management, storage, and archiving exposes the Institute to significant risk, particularly in relation to the security of sensitive personal information. This distributed environment also creates a situation where infrastructure for storing and sharing data is not equitably available to all DLCs. Central administrative units, which might be positioned to offer equitable support and services, are not always resourced at the levels needed to meet the demand. The diagram on p. 29 conveys research categories and activities that occur in different projects or that a single project might encompass.

The lifecycle of data management falls roughly into four categories: data acquisition and procurement, data preparation, data storage and sharing, and data archiving. The current state of data management at MIT presents unique challenges and opportunities in each of these four categories, and different DLCs are affected in different ways throughout the lifecycle of data management.

<u>Data Acquisition and Procurement</u>

- No support or clear and consistent Institute policies are in place for issues related to data use agreements and security requirements, particularly related to private data.
- MIT Libraries has DLC-specific legal and administrative teams actively supporting acquisition of licensed/purchased datasets, but the team size is small and service availability is not necessarily well-known. This results in an Institute-wide

distributed process to obtain for-pay datasets that leads to inefficiencies and overspending. By contrast, a well-resourced procurement service operating at the scale that the SCoC requires would be able to resolve many of these issues.

- MIT has no Institute-wide transparency regarding existing datasets, and it exercises little-to-no enforcement of security protocols at the lab or researcher level. These conditions introduce serious risks when managing sensitive or private information. Our working group also found inconsistent practices around privacy and fair use data. Some groups with FERPA data are told they have to create air-gap facilities. Some groups publishing data are told they are violating fair use. Many peer universities have more consistent policies that enable users to access private data securely without air gaps and to share information without having to navigate confusing legal standards.
- Another source of risk is data collection from the internet by faculty and students. MIT has no mechanism to communicate basic standards or to ensure that such collection is done legally.
- Data often is acquired without an initial business analysis that would reduce the risk of acquiring sensitive identifying information. For example, sensitive identifying information should not be allowed to enter the MIT environment unless it is absolutely necessary for research outcomes.
- MIT's IPIA agreement is outdated with respect to rapid and open software publication.

<u>Data Preparation</u>

- MIT Libraries has a data management services team focused on consultative support for development of data management plans and support throughout the research lifecycle. Unfortunately, the team size is small and service availability is not necessarily well-known.
- Data preparation activities are performed by individual faculty without centralized supporting infrastructure and tools and without the adoption of existing standards for describing and structuring data.
- Few centralized efforts exist to create curated and integrated datasets.
- No standard methodologies are in place for understanding and mitigating bias in datasets, although some research teams at the Institute are actively studying the

effects of bias in datasets and methods for mitigating bias.

Data Storage and Sharing

- Data is not stored and made accessible with an eye towards openness and reusability, a lost opportunity for reusing datasets.
- Currently, MIT lacks the ability to manage access and permissions so that multiple faculty and students can access the same datasets with varying permission levels.
- No data management system exists to make sure that data is renewed or discarded.
- No clear and transparent strategy or policy exists regarding data-sharing with third parties.
- MIT has no data registry, which results in redundant acquisitions of datasets and an inability to fully leverage and reuse data once acquired. The lack of documentation means that original context and purpose for datasets are not evident.
- Data management for collaborative projects is not broadly available at MIT. Increasing availability would require development of workflows as well as storage and access infrastructure, particularly for restricted data.

Data Archiving

- Mature methodologies exist for archiving smaller-scale data products, but infrastructure resources are lacking for archiving data at MIT-wide scale. The Institute must develop methodologies for archiving massive- or extreme-scale data.
- Data products are typically structured or prepared for immediate research needs, but not necessarily with an eye towards long-term archival and reuse.
- Data products are widely distributed and not well-documented after research phases ends. Storage locations are often inaccessible, obscured, or unknown, rendering archival nearly impossible.
- Connections between related products (data, publications, grants, etc.) are often lost during research phases owing to minimal metadata, creating a significant amount of post-research work to re-establish connections.

Funding/Resource Models

Although MIT relies almost exclusively on grant funding, support for research computing within the broader university environment can be seen as falling into four major buckets. Most successful research computing endeavors are supported through a balanced combination of these four buckets:

- Grant funding (including startup packages) to an individual, a department, etc.
- Donations of hardware and software.
- MIT direct-line funding—endowed fund interest or other unrestricted institutional income (licensing, fees).
- Indirect cost pool (department/lab level/university level).

Grants

- Grant-based funding can support both equipment and staff.
- Typically, grant-based funding is tied to specific objectives, and creating a rounded endeavor solely from grant funding makes it challenging to address needs that don't align with grant goals.
- Grant-funded equipment is usually purchased as a capital item. Capital purchases are not included in the pool of costs taxed to meet indirect expenses and are not subject to F&A charges.
- Staff and services costs supported through grant funding are included in the indirect-cost supporting pool and are subject to F&A. This currently encompasses grant funds used for cloud costs and also grant funding used to purchase internal services. This includes increments of storage that are below the capital-purchase threshold as well as extra support, training, and technical assistance beyond any base service.
- In some cases, grant funding can come with stipulations that complicate integration into a larger pool of resources. For example, some sponsors require that all equipment purchased under the grant be destroyed or returned at the end of the grant period. This can prohibit the use of such funds to buy partial contributions to a larger system.
- The inclusion of cloud computing costs in tax-bearing pools adds a sizable extra cost on top of an already relatively uncompetitive total cost. This creates an extra disincentive for PIs to explore cloud computing. Other institutions, such as the

University of Washington, allow PIs to apply for an F&A waiver on cloud computing[1].

- Relying very heavily on grant funding for equipment complicates retirement of space and/or power-inefficient resources. In general, participants in a grant-funded system have a reasonable perspective on retiring equipment that has outlived its usefulness, but there can be challenges to incentivizing groups to replace functioning older equipment. At present, this is a particular challenge for storage technology that has become increasingly compact.
- Many PIs are comfortable with models in which grant funds are used to buy into computing and storage that are available for their work when needed but can be shared when idle. In this model, basic system operations and base storage are provided for buy-in participants.
- An approach that is heavily or almost exclusively biased toward buy-in funding has challenges, including the provision of adequate training resources and support for core common infrastructure such as high-speed networks and high-speed storage.

Donations

- Developing collaborative partnerships to create computing facilities that benefit both parties was a key piece of both the Project Mac/Multics era and the Athena era.
- Developing strong partnerships in a coherent manner can be valuable to all involved.
- Ideally, this approach would not lead to a haves-and-have-nots situation where resources were available to a limited group.
- Direct hardware and cloud credit donations can be some of the most cost-effective ways to meet needs, but they are difficult to sustain.

MIT Direct-Line Funding

- A theme of PI survey responses and interviews was that most other major research universities have decided to make a large commitment of existing unrestricted funds or have aggressively solicited development gift funds to support research computing. Not surprisingly, many PIs argue that MIT's current model could be

[1] https://itconnect.uw.edu/research/waiver/

changed in this direction, which would positively impact research and education. We believe that alumni and friends of MIT would be interested in helping support a development effort to create a more comprehensive and integrated Institute-wide research computing infrastructure ecosystem. In that scenario, funds do not have to be part of a zero-sum game that reduces funds available elsewhere. At present, fundraising efforts in this area do not appear as fully organized, focused, and coherent as they could be. It may be time to start viewing the creation and sustaining of cyberinfrastructure on par in stature and value with the creation of physical building infrastructure.

Indirect Cost Pool

- A final component source for funding of the research computing infrastructure for the SCoC and MIT as a whole is taxation, either at the laboratory or Institute level. Current efforts at encouraging and developing larger group and Institute-wide services have reduced significantly tax burdens associated with campus facility and power costs. At present, no mechanism exists to leverage these savings in a manner that directly benefits research computing efforts. Adjusting practices around indirect cost recovery is a very complicated issue, but some precedent exists.

Pros and Cons of Funding Models

The balance between the use of different funding models and cost-recovery practices impacts behavior and shapes the landscape. Adjusting funding models takes time and is a delicate operation. Shifting costs into the indirect cost pool can reduce direct charges to PIs and help steer service models in a specific direction. Such shifts, however, also put pressure on the institutional or laboratory tax rate. Indirect cost pool recovery is also constrained by federal regulations, which can create uncertainty and risk over what is reasonable to undertake. Commercial cloud providers have long been lobbying executive levels of the federal government to update guidance to explicitly exempt cloud services from tax. This could reduce a disincentive to the direct purchasing of cloud cycles. Whether this would be sufficient to increase use is unclear.

Very heavy reliance on direct grant funding alone tends to encourage structures in which each lab works independently. Grant funders are reluctant to fund large general-

purpose resources unless there is some degree of institutional matching. This makes it hard to grow very large scale systems through this model alone.

Equipment and cloud credit donations through central resource development do help to provide valuable services. These resources can be very effective, but they are also somewhat hard to plan around. They almost inevitably require some extra investment to fully leverage, and they are not necessarily sustainable. Donations to individual labs or researchers in the name of MIT have historically not been widely shared.

Long-term guaranteed support from general funds, which is not unusual elsewhere, runs somewhat counter to MIT culture. It often involves difficult, effectively zero-sum tradeoffs among many priorities, and it can be quite tricky to resolve different perspectives. Some campus interviews revealed that members of the community were highly skeptical that it was in the nature of MIT to heavily support any central computational infrastructure. However, this funding model is the only truly sustainable approach.

POTENTIAL MODELS

This section broadly describes the different pathways that are available for computing infrastructure. This report does not advocate specifically for any of these models, but provides pros and cons that should be considered. Typically, no institution selects a single one of these models but instead relies on a hybrid approach where the debate is centered on the distribution among multiple options.

Completely Centralized

- Deploy one large centralized facility. Use all MIT available space at MGHPCC for one large shared computing infrastructure.
- Set up large data storage with various levels of access and licensing capability as well as high-bandwidth communication with campus for transferring large datasets or experimental results.
- Hire sufficient staff at both sites to maintain and facilitate use.

- Develop a good priority system to balance requirements and needs equitability by providing access to anyone with an MIT account.
- Explore the potential for partnership with several area institutions to reach multiple megawatts of capability and manage operating costs. Potential examples include launching a regional partnership with other MGHPCC members or entering into a partnership with Lincoln Lab.
- Provide training to all areas of MIT and flexible software to address all needs.

<u>Pros of a Completely Centralized Model</u>

- Simpler control of access and security.
    - Applies to campus software licenses and data agreements.
    - Lowers risk to the Institute.
    - Simplifies equitable distribution of resources.
- Strong possibilities of building a very large system though partnership with regional partners.
- Could provide seemingly infinite computing access to groups that are typically resource-limited.
- Could energize new and vastly more ambitious research thinking and possibilities in many new domains.
- Some infrastructure is already in place at MGHPCC, and most participants are quite satisfied.

<u>Cons of a Completely Centralized Model</u>

- Requires a major cultural shift at MIT.
    - Many heavy users favor a decentralized approach because they are well funded and connected.
    - Definitely would garner push-back from MIT cloud users (and possibly from cloud providers) and elicit a fair amount of skepticism and doubt at MIT.
- Lacks flexibility.
    - Would likely lead to regrowth of systems on campus if software and operations were not flexible enough.
- Does not meet requirements for everyone without a complex software setup.
    - Some data licenses might be too restrictive for such systems.

    - Export control software couldn't be supported.
- Long queues.
    - Most major computing infrastructures suffer from large turnaround times.
- High cost.
    - Maintenance obligations.
    - Requires sustainability plan.
    - To really be transformative, it would require a substantial, sustained budget.
    - Would require a decade-long commitment and budget for support, maintenance, and proactive training to get the most out of it.
- Requires vigilance not to do things that would be cheaper and better done elsewhere.

Partially Centralized but Managed within DLCs

- Large centralized system at MGHPCC <50 percent of the total space.
- Large data storage with various levels of access and licensing capability.
- Dedicated line-item funding with sustainability plan.
    - Hardware replacement on centralized system.
    - Clearing rack space after certain number of years.
    - Support staff budget for centralized resource, remote hardware services, integrating centralized system with willing departmental scale resources (e.g. Physics, Brain and Cog), and cloud.
    - Common storage platform available to all systems, including common/unifying authentication (technology for this now exists and is increasingly being deployed).
    - Transparent long-term roadmap reflecting commitment.
- Distributed workload queue system that can accommodate many use cases.
- Rack space available for additional specialized needs (at MGHPCC, Bates, and OC11) that can potentially integrate with central service.
- Integrate with cloud resources and with available cloud credits for reducing queue constraints and for those many users requiring only single node jobs.
- Solid technical training and support available for using both MIT resources and cloud counterparts, potentially with unified portal.

- Cloud credits or subsidized/bulk cloud discounts with some sort of data preservation gap coverage.

Pros of a Partially Centralized Model

- Provides many of the pros of the centralized option.
- Meets diverse needs of population.
- Models a scalable facilities approach to core MIT infrastructure (e.g., offices, classrooms, bathrooms).
- Can evolve and fully integrate with existing large-resource partnerships at Lincoln Lab, Bates, other regional universities, and elsewhere.
- Executed well, it will create a whole that is a lot more than the sum of its parts by harnessing energy and ideas all across campus regarding support, software, physical hardware, cloud partnering, and so on.
- Lots of flexibility for engaging the student population in various areas.

Cons of a Partially Centralized Model

- Requires financial sustainability plan.
    - Centralized portion would require MIT line-item support.
- Support staff would be expected to support centralized and decentralized systems.
    - Scope creep in their jobs.
    - Less equitable support, since needs of dominant groups could overwhelm staff.
- Requires physical sustainability plan to free up space.
    - Flexible space would be dominated by a few with no incentive to free up space (first-come-first-serve model, fee model?).
    - Any retirement model might increase probability of systems coming back to campus.
- Increased risk.
    - Cyber and licensing violations increase with the number of unique systems. Audit compliance would be necessary.
- Depending on level of decentralization of support costs, it could be challenging to create a uniform quality support and training/on-boarding experience.

Fully Decentralized

- This approach is essentially the current operational model in which every PI is self-sufficient and certain centralization decisions are made at the DLC level.
- This approach is applicable to both cloud or physical hardware or to data.
- Support is typically provided by the PI or DLCs that can afford support.

<u>Pros of a Fully Decentralized Model</u>

- Meets precise needs of some PIs with the means to purchase hardware.
- Very flexible from a research point of view (only limited by the funds/donations you can raise).
- Creates a sense of resourcefulness.
- Risk limited to individual groups/PIs.
- Decentralizes the risks to individual systems.

<u>Cons of a Fully Decentralized Model</u>

- Barrier of entry is costly to non-specialist and certain scales of computation are inaccessible.
- Cybersecurity issues.
    - Huge risk since PIs are largely responsible for their own updates, security patches, and access restriction. Many systems use software that is obsolete and no longer supported. Many non-experts manage systems.
- Data licensing and access becomes a major issue.
    - Difficult to share across groups.
    - Every new license suffers from lack of institutional memory.
    - Every group must reinvent the wheel in meeting licensing restrictions and requirements.
    - Huge exposure risk in case of breach (lack of system update, maintenance, proper security follow-ups, etc.).
- Difficult to maintain.
- Poor educational integration.
- Sustainability is a major issue.

    - Many government funding agencies are becoming stricter in allowing for hardware purchases.
    - No incentive to remove old hardware.
- Poor space and cost management.
    - Constantly facing lack of space and inefficient use of space/power/cooling balances.
- Inequitable (could leave large parts of MIT on the outside).
    - System provides no equity in computing access and rack space (first-come-first-serve basis).
    - Classroom instruction often leverages instructor research resources. Unequal course experience depending on instructor.
- For most PIs, the absolute costs of this approach mean their research is more constrained than necessary.

Predominantly Cloud-Based

- Cloud providers have become more capable of providing a range of services that can well serve the research community. Good migration capability is necessary to avoid being locked in with a single provider.

<u>Pros of an Entirely Cloud-Based Model</u>

- Scalable in minutes, and you only pay for what you use.
    - No queue.
    - Low entry costs (effectively zero).
    - Access to very diverse resources (from low cost CPUs to modern GPUs).
    - Leverages the commercial cloud's amazing self-service capabilities, which allow capable and well-funded groups to move fast and innovate at a great pace.
- Highly adaptable.
    - Very well documented and very flexible.
- No use of campus space.
- Easier to integrate with education.
- Large number of services via APIs (e.g., Tensorflow, Spark, Kubernetes).

- Highly developed security and ability to manage centralized policies.
- Centralized billing, control, and some economies of scale.
- No fundamental need for central investment or support beyond network.

Cons of an Entirely Cloud-Based Model

- A modest number of groups need hands-on access to cluster hardware for performance research. This was raised in the all-PI research computing surveys.
- High cost for long-term such that sustained, truly 24x7 computation would cripple some computationally intensive research efforts.
    - Time and cost to transfer large data sets, as with all off-campus sites, could be a factor.
    - Cost for large storage is high and projects with large storage needs could be stopped in their tracks by budget challenges.
- Export control or data licensing issues.
    - Many tools, software, and datasets cannot be stored on a remote cloud. This is highly dependent on the field of research, but these conditions do exist on campus.
    - Breaks various existing data-use agreements that often require raw data to stay "within MIT." Although these could be renegotiated in theory, it is difficult in practice.
- Steep learning curve, not easy to use.
    - Every cloud provider is different. Many options would require lots of help to get set up and functional. Any transitions from provider to provider would be painful for many.
- No appropriate MIT funding model (i.e., overhead costs on research contracts).
    - Data egress charges can be surprisingly high.
    - Researchers could end up losing all their data when funds run out without some potentially costly institutional backstop.
- Exclusive cloud has the potential to incentivize some PIs to switch back to building out local campus resources or to create physical facilities tensions.
- Difficult to manage spending limits.
    - Better tools are needed to facilitate setting spending limits, distributing credits, and providing proper access control.

- Evidence suggests that this would cost more in absolute dollars, although cost burdens could shift from central-facilities costs to direct-PI costs.
- Occasionally, market/commercial conditions can impact research. For example, the current emphasis on deep networks in machine learning creates a significant disadvantage for academic researchers relative to their industrial counterparts that have access to orders of magnitude in computing and training data.
- MIT doesn't even entirely outsource its electricity supply. In fact, the Institute is currently building two 22MW generator sets on Albany Street. These are being built in part to ensure resilience with respect to commercial supplies and changes in market dynamics.
- If supported through credit gifts or SCoC funding, the SCoC could become an island of services only available to a subset of the community. This issue was raised in many interviews.
- Courts can subpoena the cloud provider and not MIT for material.

ETHICAL CONSIDERATIONS

Our working group identified the following infrastructure-specific ethical considerations.

Climate Impact

Collectively, the world's data centers are the fourth largest consumer of electricity after the US, China, and the EU. It is essential that computing infrastructure consider CO2 emissions. MIT should hold our cloud providers to the standard we have set at MGHPCC. The good news is that most cloud providers have plans to reduce CO2. When MIT is considering cloud purchases/gifts, it should encourage cloud providers to keep their CO2 promises.

Our Own Environmental Impact

MIT's major computing facilities in Holyoke are nearly 98 percent carbon-emissions free. The savings in avoided CO2 emissions from using this facility are comparable to all the on-campus saving from building and facilities upgrades. The Holyoke facilities also were developed on a remediated brownfield industrial site in contrast to the common approach

of developing and paving over pristine arable land. Continuing to uphold these sorts of practices going forward would be a laudable goal.

Community Impact

MIT's efforts to provide computing infrastructure to the Holyoke and Cambridge communities should be encouraged. MIT should be cognizant of the community impacts of cloud providers and encourage them to sustain their positive impacts and minimize negative impacts (e.g., not providing living wages or affordable health insurance). MIT currently works with Holyoke to create internship opportunities for community college students that provide economic opportunities that would not otherwise be available.

Privacy/Trust/Anonymity/Anonymization/Authentication/Authorization

The MIT-operated computing infrastructure is based on a model where infrastructure access is granted by accountable humans and anonymous usage is prevented. This approach should continue to guide MIT even as we make it easier for work to move seamlessly around our hybrid clouds. MIT computing infrastructure personnel are often the architects and implementers of the technical means for upholding data protection agreements of many types (educational, corporate, government). MIT should be aware of how data is being used by cloud providers and make sure students can opt out. MIT should be aware of how the digital footprint of students from other countries is being monitored by their home countries so these students are able to fully participate in MIT discourse.

Open ResearchWare

Create open availability for data and software portal-related activities that people are engaged in to make research artifacts (i.e., data and tools beyond papers). As with OpenCourseWare and open edX but with lower visibility, a few groups provide services globally from labs that consider this a valuable service.

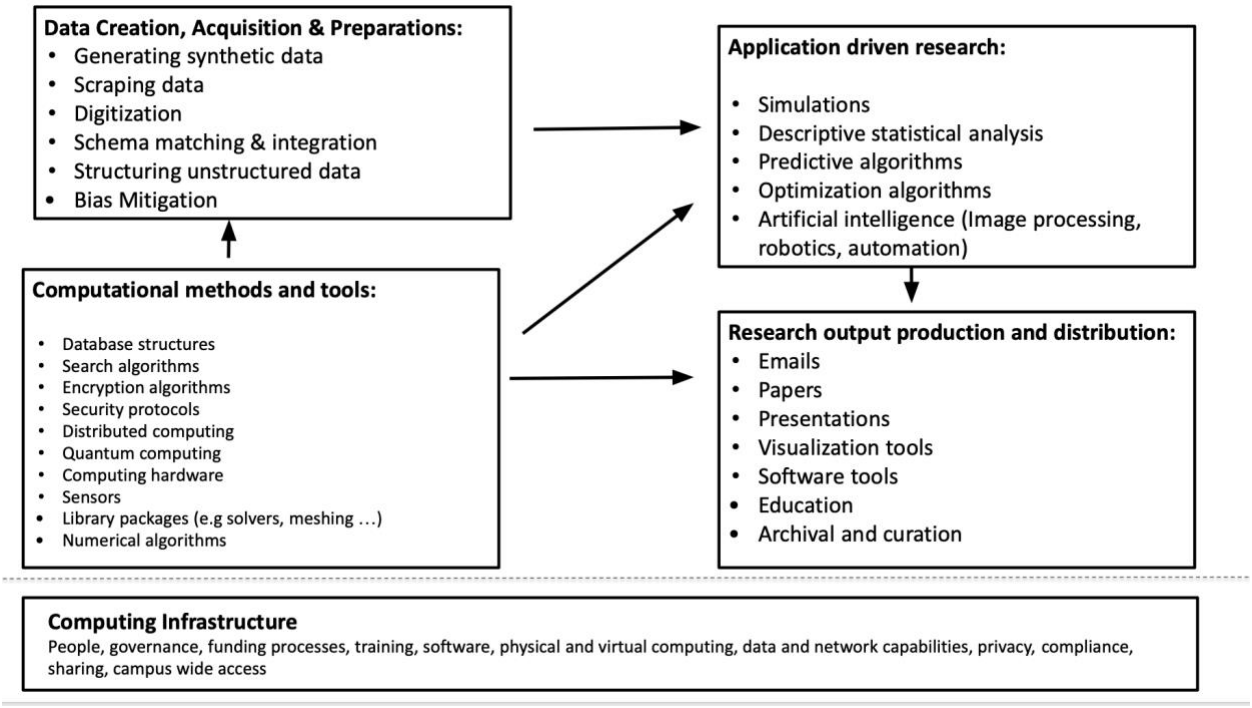
Equitable Access – Measuring Access

MIT must have a means of determining whether needs are met. How do we measure equitable access to infrastructure? How do we ensure that all members of the community are getting equal access to resources, computing time, and relevant support and tools?

INSTITUTIONAL ANALYSIS

THE PRESENT STATE OF RESOURCES

The diagram below highlights the many ways in which the computing infrastructure underpins the full span of research and education activities at MIT and conveys the richness of different activities and perspectives that expressed the need for computational infrastructure in some form.



The diagram conveys research categories and activities that occur in different projects or that a single project might encompass. Different participants in MIT research are sometimes more strongly tied to one category or another, while other research community members may move between categories and activities very dynamically. All the categories have needs for computing infrastructure of some sort, although the needs may look different depending on which category or activity lens is being used. The arrows convey a general downstream flow of research, but ideas and discoveries often feed back in the opposite direction to the arrows, with “downstream” research seeding or becoming research in an “upstream” area.

One goal of the diagram is that everyone involved in research at MIT should be able to identify one or more locations that correspond to their research and education activities and computational needs. The bottom of the diagram represents our working group's vision for MIT's research computing infrastructure. It is a separate non-research activity, but as indicated by the dotted horizontal line, there is semi-permeable membrane quality to the interface between the infrastructure (which includes people and budgets as well as computing/data/network resources) and the research community.

Commercial Software

IS&T centrally manages ~300 software packages that are available to faculty and students. A number of software vendors are trying to switch to a pay-as-you-go model with metered licensing tied to cloud services. This is causing disruption to existing processes of paying for software use. The situation presents IS&T with some practical challenges in how to continue to affordably deliver services that the broad MIT community has come to appreciate and rely on. In general, these metered models seem to be good for software providers but not as obviously beneficial to customers. In the realm of academic publications, adoption of somewhat similar models for journal charging has created severe pressure on library budgets and is causing challenges for all university systems (https://www.universityofcalifornia.edu/news/why-uc-split-publishing-giant-elsevier). At present, the MIT software landscape has not been that affected by the move toward this sort of pay-as-you-go licensing. Looking forward, this could become a challenge.

Cloud Computing Use at MIT

Approximate spending on the four major cloud computing vendors is listed below. These costs reflect purchases made through MIT's purchase order system, via an MIT P-Card or through vendor credits.

- Amazon Web Services (AWS).
    - 2,300 accounts across MIT.
    - MIT's current spending is $7.5M annually ($3M of this by IS&T for virtual machine infrastructure).
    - AWS also provides credits for cloud usage to PIs and programs.

    - The AWS Educate program can provide $200 credit to students who apply.
- Google Cloud Platform.
    - ~75 accounts at MIT.
    - Current spending is ~$200K annually.
    - Over $1M in credits for research programs.
- IBM Cloud Platform
    - Used heavily by MIT–IBM Watson AI Lab to support MIT Quest–Core.
    - Growing use in MIT Quest–Bridge.
    - Total annual spending ~$600K.
- Microsoft Azure.
    - $100K in annual spending.
    - Microsoft provides credits to PIs and research programs.

PRESENT NEEDS & PAIN POINTS

Our working group conducted question and answer sessions with representatives from all DLCs who were willing to meet with members of our group. These use cases try to summarize and synthesize some of the information we gathered. One theme observed by many is that not only are people not aware of all the resources at MIT, they also don't necessarily know that they don't know. Numerous examples were encountered during interviews and town halls of people who broadly knew of existing services in data and computing but instinctively assumed these services were not meant for, or could not be relied upon to meet the needs of, their particular group or lab. Others knew, and had not found, adequate support to sustain their projects.

Literacy and Training

The level of computing literacy needed to function successfully in education and research is increasing. Today, almost everyone requires a laptop computer with access to email and other collaborative tools to function in any business environment. While most have basic skills in managing documents on a Mac or PC, the level of literacy among faculty, staff, and students varies widely. The level of confidence varies just as widely, with some community members fearless about experimenting with new technologies and others hesitant to jump in.

Students

The route to literacy for students is usually via the required courses for a variety of degrees, but it should be noted that many students come to MIT well-prepared. As more departments develop minors in CS or joint programs with CS, the route for students is through the CS department (e.g., 6.0001+6.0002 Introduction to Computer Science, 6.0004 Computation Structures, 6.009 Fundamentals of Programming).

Post-Doctoral Staff / Graduate Students

Faculty in most DLCs tend to hand off the details of maintaining servers and provisioning and developing software to their post-docs or graduate students. In polling post-doctorate staff and graduate students, one of their most frequently articulated requests is for training in this skill set.

Faculty / Research Staff

The standard of computer literacy among faculty and research staff varies widely, and we have no available metrics to define it. MIT has no well-designed route for members of these groups to achieve literacy, even though the necessity for faculty to be knowledgeable is certainly increasing.

Conventional HPC

Conventional HPC users seek large computing environments for high throughput simulations with fast interconnect. They also require large nearby storage for results and necessary databases. Typical HPC users are currently not well set up for running on newer hardware (e.g., GPU) and they often use highly specialized software with highly varying requirements, libraries, and support. Some software is licensed-based but most is open source and community-developed or in-house tools.

Many Lightweight Jobs / CSAIL

Computer science researchers often require access to large computing resources to demonstrate their research at relevant scales. These researchers often desire to test at

scale and publish comparisons among hardware and cloud resources. Hardware and software diversity can also be an important part of this research.

- Systems ideally should be developed such that they are 1) interactive and 2) valuable to both those who have computational needs and those engaged in computer science.
- Presently, it is very time consuming for students to set up accounts in order to conduct projects as part of courses. It would be beneficial to have a common API for logging onto various systems to make it easy for those who already have an Athena account to get onto those systems.
- Today, projects often need to retool themselves in order to be able to use various donated computing resources (such as IBM and AWS credits). Solutions that could reduce or eliminate the need for such retooling would be very valuable.
- The move towards programmable infrastructure in the cloud (creating and provisioning machines using languages such as Python, TypeScript, Go) is rapidly becoming available. Once these programs have been developed, complex pipelines can be defined and launched using a single command. The Broad Institute has used this strategy to enable researchers to focus on their research goal without having to learn the details of cloud computing. https://www.broadinstitute.org/news/8065

Protected Data Users

The needs of protected data users at MIT fall along a spectrum, and infrastructure solutions to meet these needs are informed by the sensitivity of the data, the modality or format of the data, the type of access to the data required by researchers (including collaborative access), and the complexity of the data. Dedicated Institute-wide investigation is needed in order to fully develop a set of use cases that represent the breadth and depth of data usage at MIT.

Cross-Cutting Themes and Issues

- Datasets sensitivity. Datasets obtained by faculty and by MIT could have varying requirements of storage and data management. On the high end, some data requires stand-alone and potentially physically isolated computing infrastructure (e.g.,

personal information of students or data received from the FDA). The next level is data that can be handled in a centralized infrastructure but with very stringent access and communication constraints (e.g., healthcare data). A third level simply requires protection under NDA. Clearly, a continuum of scenarios exists. The requirements can be driven by laws and regulations or imposed by the data provider, and the level of individual liability of faculty could also vary. As a result, MIT must develop appropriate infrastructure and related processes to support these different scenarios.

- Data modalities. This is another area where there is a continuum of scenarios. Some modalities such as images, other analog signals, and DNA data can require a tremendous amount of storage and memory. Other modalities, such as text and numeric data, require a much lower level of storage per unit of data but might be obtained in extremely large quantities. MIT should expect to encounter an increasing percentage of datasets with integrated modalities (e.g., healthcare and manufacturing, the arts). Another expected trend would be increasing the volume of efforts to digitize data, particularly in fields where raw data is typically in analog or even non-electronic form.
- Data access and sharing. The current dominant paradigm is that data is managed by individual or small groups of faculty. However, MIT already has and is expected to have an increasing number of initiatives that will attempt to enable large-scale research programs through the acquisition of large datasets, some of which will be sensitive (ICU data at IMES is a current example of de-identified data). Such programs will require a very different infrastructure and support system to manage permissions, ensure data security, and enable data and research output sharing. Another potential scenario regarding access would be to create encryption mechanisms that allow data manipulation with partial data access.
- Data acquisition. Many individual faculty and entities at MIT regularly interact with companies and organizations to obtain private and commercial datasets at different level of cleanliness. Another aspect of this is active data collection through scraping, sensors, and experiments where each mode requires different infrastructure.
- Data connectivity and search. Many datasets have layers, each representing a different level of data manipulation. Research tasks require the ability to move back

and forth throughout the layers. The ability to search large datasets at scale is also necessary.

- Data for teaching. An increasing number of courses could benefit from in-class and out-of-class exercises and other educational activities that are enabled by real and synthetic datasets. These must be stored and accessed in a manner that enables work in the classroom, lab, or home.

Humanities and Educational Users

Although some in the humanities, arts, and social sciences have more specialized research projects and needs, we have paired humanities and education because much of computing infrastructure for each is centered around two aspects: 1) teaching with computing (tools—including hybrid classrooms and software that enables teaching as well as student work), and various student information systems, such as an LMS, and 2) exposing key audiences (faculty, students, teaching staff) to computing and how it might impact their teaching, research, and work more broadly.

More than access to technology/enabling technologies per se or infinite computing, the most sought-after resources seem to be expertise, training, and in particular, equity in terms of who is able to gain access to resources. This includes material resources such as suitable classroom spaces and funding streams as well as human resources such as early-career, tech-savvy collaborators and programmers. Having sustained, reliable, constantly renewing computing infrastructure to support teaching, learning, and research is also essential (i.e., faculty and students cannot be responsible for maintenance and self-service without robust guidance). In addition, infrastructure must be embedded in residential learning, not incidental or tied to hybrid and cost-recovery-dependent models.

All stakeholders we spoke with were adamant that computing resources and expertise be truly infused into teaching and learning from the ground up. They emphasized that while central and top-down support will be needed, past models of creating an external innovation engine or an entity that works outside the normal governance processes (even with the worthy goal of spurring innovation) will not be sufficient and more likely could be detrimental.

In speaking with various stakeholders, we learned that having access to a robust computing infrastructure will provide numerous benefits. Such access also dovetails with strategic priorities for residential teaching and learning identified in earlier faculty-led task forces. Those include the task forces on the Undergraduate Educational Commons and on the Future of MIT Education, as well as task forces launched by the Office of the Dean for Undergraduate Education and the more recently created Office of the Vice Chancellor.

Goals include:

- Provide students, faculty, and staff with the best tools and classrooms to deliver the optimal student academic experience.
- Provide more effective navigation for students to MIT resources, including both digital and in-person components.
- Provide students with sufficient research computing resources to best support their needs on campus, such as centralized resources or cloud access.
- Define and provide tools/technologies that best support open and professional learning for students and those who serve them.
- Hire new/additional learning engineering staff and disciplinarily fluent postdoctoral research associates to support faculty development of innovative pedagogy, use of new systems, and so on.
- Incentivize faculty to collaborate in education across departments with the provision of teaching resources. It should support projects and experiments that map connections in topics and outcomes across the curriculum, employ online resources to facilitate connections, and exploit opportunities to use modular approaches to increase flexibility. These types of experiments will facilitate studies of the benefits of connecting content in new ways.

In addition to spurring access to more robust computational resources, the SCoC provides a catalytic moment to think more broadly about what is needed to best support teaching and learning at MIT. This might entail:

- New kinds of learning systems such as Canvas (including assessments).
- New student information systems/learning data management.
- Student learning tools (e.g., iPads for all, Oculus for everyone).

- Virtual labs/simulators.
- Novel classrooms (TEAL+, virtual, etc.).
- Student-based cloud storage/enterprise systems.
- Student access to research computing facilities.
- Set number of FTE for learning engineers, support staff.
- Training for faculty, staff, students.
- Dedicated non-ladder, non-tenured teaching and research associates.

Potential Needs by Audience

Faculty/Teaching Staff

Thinking of computational infrastructure in a general sense, the SCoC provides an opportunity to better infuse residential learning with the benefits of computing. Success will be underpinned by an integrative approach, and most stakeholders indicated that "we cannot teach students about the future of computing when we have classrooms, courses, curricula, and student systems that are rooted in the 19th century."

Most important, the SCoC will provide an opportunity to consider the proper incentive structure needed to inspire and support innovative pedagogy that takes advantage of computation, both as a teaching tool and as a way to enhance disciplines (e.g., digital humanities, data-driven political science).

Undergraduate and Graduate Students

Students view the SCoC as a platform that will prepare them for the future of computing and the future of computing-informed fields. It is important to note, students said, that there are those who are majoring in computing disciplines or have expertise due to their research or lab affiliation and those who might be eager to learn but lack expertise and/or an obvious way to engage.

Beyond curricular innovations, enabling infrastructure could include:

- Support for tools/mentorship/training.
- Student-focused resources/a lab environment for computing.
- Clear ground rules/equity of access.

- Defining top-down goals to influence strategy.

Academic Departments/DLCs

The academic departments/DLCs are an implied audience, as they provide the bulk of the teaching, research, and other activities (including and beyond the General Institute Requirements for undergraduates and outside some aspects of graduate education and student living). MIT's 50+ departments are responsible for various GIRs, courses, degree programs, and in many cases, the pedagogical and technical infrastructure that supports their faculty and students.

While MIT's five Schools provide some overarching support and infrastructure (i.e.., money and staffing), it is an open question as to how the creation of the SCoC might centralize certain resources and/or replicate models akin to MITx fellows or internal grant-based funding (such as the D'Arbeloff awards or MacVicar fellowships).

The sheer diversity of how computing resources are used to enhance teaching is not surprising given the nature of MIT, as is the variance about what a particular department may need now and in the future.

Comments from public forums

At public forums held so far, discussion reiterated many of the themes that arose elsewhere, including:

- The increasing role of sensitive data (medical, financial, etc.) in interesting research and the desire to have a streamlined process for data-use agreements, managing access, carrying out research while complying with restrictions.
- Services designed to help navigate data-use agreements actually exist on campus but are not widely visible.
- Digital humanities often has video, audio, and text processing needs that require special services, and it is not clear that those services are being thought about.
- Facilities such as MIT.nano are built without any high-speed networking, even though they are becoming locations with multiple, relatively high-bit rate devices

(such as CryoEM machines) that require bursts of high-speed data transfer for analysis and possible digital archive.

- An approach that leverages the interest of the student population in really creating a unifying framework that could be cost-effective and mutually beneficial.
- Services and research software engineer expertise for supporting reproducible digital research are relatively piecemeal at the moment. This seems counter to the credible research mission of MIT.
- There may be opportunities around data sharing and research tool sharing that research groups would be interested in rallying around.
- Growing support and training services that span the major physical and cloud resources available would be very welcome.

Comments from the Web Ideas Bank

So far the working group has received four comments. The first comment pointed out the value of research software-engineer positions as a way to provide training and support. The other three comments discussed models for delivering computing and storage:

- One proposed a sustained set of services centrally supported and modeled after FAS.
- A second proposed a decentralized approach, but with some central funding to help decentralized groups.
- The third argued for avoiding a one-size-fits-all single central resource and expressed interest in training and support for groups wanting to build their own systems.

At first glance, those comments seem to reflect a slight "Man, Boy, Donkey" dilemma (https://en.wikipedia.org/wiki/The_miller,_his_son_and_the_donkey) for our working group. On the whole, however, the comments are consistent with some sustained central activity that provides help to various models and potentially creates a more capable whole spanning a range of approaches.

COMPARISON TO PEER INSTITUTIONS

Computing infrastructure at peer institutions—Harvard (https://www.rc.fas.harvard.edu), Yale (https://research.computing.yale.edu), Princeton

(https://researchcomputing.princeton.edu), Stanford (https://srcc.stanford.edu) and Berkeley (http://research-it.berkeley.edu/programs/berkeley-research-computing)—came up as a topic in comments and in interviews. In general, MIT has a less robust central financial commitment to computing infrastructure resources. As a result, the MIT ecosystem generates more self-organized activity. At all the peer universities we highlighted, the central groups include multiple research software-engineering resources that provide training and support services in different domains. This is in addition to core-services aspects of the groups that support software and hardware infrastructure, data management services and consultation, and archival and preservation of the products of research.

These organizations also fund IT staff activities to provide seconded services around specialized networking for low-latency networking and parallel storage systems. Separate, additional budgets are also generally managed for data curation support and cybersecurity support. In general, these peer organizations have come to terms with the notion that research computing infrastructure is becoming more like other core campus infrastructure entities such as buildings, plumbing, and electricity, despite having the word "research" attached. That perspective eases budgeting for a focused, sustained commitment to an enterprise that manages the services. In these organizations, the "research computing" phrase implies services delivered by technical groups that have the right research background but that are not undertaking research. In many cases, reaching that understanding is a result of doing the right thing after exhausting all other available options.

Organizationally, we observed a universal approach of reflecting an alignment with research missions by clearly distinguishing research computing efforts from enterprise IT activities. Typically, this entails direct organizational hosting within the academic provost-led side of the university, albeit with strong links to IT leadership. In some cases, there is a 50:50 split between IT reporting and provost reporting. In other cases, reporting is 100 percent on the academic side. No organizations are solely and fully hosted in the enterprise IT component of their universities. The overall research-computing-infrastructure organizations typically do not reside within an academic department but operate as a separate entity within the academic side that collaborates with all departments. The groups work closely with all departments but are structured to be cross-disciplinary, oriented to access for all, and more operationally focused than a group housed in a single department

or school would tend to be. Governance is organized around an oversight body that has representation from all stakeholder departments and provides guidance to the infrastructure organization and helps develop and advocate for long-term budget and planning roadmaps. Senior administration have come to terms with responding to such budget proposals.

It is not surprising that all institutions struggle with achieving the right balance relative to other priorities. All involve a mix of large, universally available resources, cloud service support, and smaller departmental and individual PI activities. In some cases, the large, universally available resources have grown to be very dominant. Even in those cases, however, the overall landscape is still a hybrid of evolving technologies. Two peer institutions that were highlighted have very active partnerships with large laboratories (SLAC and Lawrence Berkeley). Those synergistic activities provide additional expertise and resources from the laboratories along the lines of MIT Lincoln collaborations, but at a larger scale.

Budgetarily, some groups are purely funded centrally, but most operate in a hybrid fashion. They have a core base budget with a long-term guarantee, but they also raise or can accept grant funds. The groups also offset base budget with for-fee services that go beyond base services and are available at no direct charge. Examples of fee services include consulting services, additional storage, additional compute cycles, and increased access services for time sensitive work. Budget line-items typically include salaries as well as funds for some hardware resources, cloud access, training activities, network resources, maintenance services and depreciation, and renewal budget.

Several people noted that the grass is not always as green as it looks at these peer enterprises. As everywhere, making decisions about budget, space, resources and collaboration is never easy, and no institution has developed an out-and-out utopia (yet!). Our working group also reached out to academic and staff colleagues at Caltech. Caltech undertook a concerted effort at steering researchers toward purchasing cloud resources as the main approach to computing. The decision-makers behind that effort have decided to adjust slightly and are currently restoring some local services. Members of our working group found few universities that had managed to develop very large and very long-term

rate discounts or cloud-credit gifts from major cloud vendors with the exception of the University of Wisconsin, which has negotiated discounted rates with Microsoft and Amazon. The closest local partnership may be between Broad Institute and Google. The two have been working closely on a general computing portal for Broad work (https://portal.firecloud.org/). This has created a semi-general, science-computing-oriented container execution engine with an associated storage model. The system offers some free credits, but as with many such initiatives, the credits appear poised to sunset within a year. Creating a new partnership with commercial cloud services may be an opportunity for senior administration resource development.

LEGACY

The MIT landscape for broad computing infrastructure has varied cyclically over many decades, depending somewhat on the level of innovation in technology and research needs. Two particularly significant major institutional efforts took place in the late 1950s and 1960s—Project Mac/Multics era partnerships with IBM, GE, the Office of Naval Research and others, and the 1980s Athena era partnerships with IBM, Digital Equipment, and others. Both efforts produced impacts beyond MIT and provided the campus with sustained infrastructure for research. Those efforts have many parallels to the landscape today, with the exceptions that:

- Today, a commercial cloud industry that first appeared in a crude form in 2006 is creating an emerging utility service that has never really existed.
- A remarkable and organic open-source ecosystem has emerged over the last 25 years, driven from the emergence of the open-source Linux operating system in 1992. The full impact of open source has become remarkably empowering and democratizing in research computing in the last five to 10 years.

Those two factors are a potential disturbance to a landscape that in many ways has been in an otherwise steady rhythm for 60 years, even as technology evolved significantly.

1950s–60s, MULTICS, AND PROJECT MAC

As far back as 1957, MIT led substantial regional initiatives that spanned physical sciences and emerging computer science areas, bringing together industry, academia, and government. The work in that era was one of the first efforts to develop the scale of multi-user, multi-institutional interactive supercomputing.

The project was led by MIT but involved collaborations with 25 New England universities and colleges. The project was supported through a mix of institutional resources, industry sponsorship (from OBM and later GE), and government partnership (through ONR). The activity included special physical plant (the TEAL laboratory 26-152 is a remaining

structure), support staff, 24 research-associate positions, and a dedicated network fabric (see map from Distributive Computer Networking: Making It Work on a Regional Basis[2]).

The 1950s–60s era work was initiated by physicists Phil Morse and Herman Feschbach and others with strong support from Provost Julius Stratton and IBM. The project evolved over a decade to be the heart of Project Mac and led to the creation of one of the earliest time-shared operating systems, Multics. The project placed interactive terminals around campus and at remote partner university locations. To support this mode of usage, numerous technical advances were made in operating-system technology.

This effort lasted for roughly 15 years and was one of the inspirations for early, standardized, wide-area, packet-switched networking that went on to become the Internet. After roughly 12 years, the computing services developed in this era transitioned to a central IT service operated alongside MIT business operations in the late 1960s. That subsequently was wound down over a period of years as mini computers grew in capabilities.

1980S, ATHENA PROJECT LESSONS LEARNED

In response to the maturing and shrinking of computers, Project Athena emerged to form the first fully autonomous but networkable deskside/desktop workstations. Its genesis came in part from a vision of computing as a platform for digital learning. Athena was supported through industry collaborations with IBM, Digital Equipment, and others. Like previous generation projects, the Athena project led to the development of new core software—in this case, for supporting distributed computing that could move with students and present a unified, graphical environment all around campus. Outgrowths of Athena include the X windows system, the first graphical widget systems for rich graphical displays, and the beginnings of the open source movement. Those technologies are the unnoticed infrastructural underpinnings of much contemporary computing. In the Athena growth area, those were truly transformative inventions.

[2] DOI: 10.1126/science.189.4202.523

Additional relevant details about Athena's infrastructure, support model, and sustainability include:

Hardware

- At its height (1991), Athena had ~1,300 workstations deployed around MIT.
- In 2019, 225 Athena workstations were deployed at MIT across 15 locations.
- Hardware has changed over the years, but it has included workstations from DEC, Sun, Silicon Graphics, Windows, Mac.

Software

- Athena introduced the concept of "lockers," a network file-system abstraction that provided shared file storage, web hosting, and software application delivery for classes (an early version of a container).
- ~1,900 "lockers" have been used on Athena to deliver computational environments for courses.
- Available software packages for students, such as MATLAB and Maple, were easily accessed without the need for installation.

Technical Support

- Athena On-Line Consulting.
    - Athena Consulting support was provided by a team of paid student consultants (at its height, >50 students) that were supported by a team of three-to-four full-time staff members.
    - Tools were custom built for the environment and, in many ways, were revolutionary for their day.
    - Student consultants had easy access to domain subject matter experts (often former consultants themselves), operations staff, and developers.
    - Real-time chat connections were established between users and on-duty consultants.
    - Equivalent system for TAs.
    - A healthy competition arose among consultants to see who could work with the most users at the same time.

- Athena Watchmakers.
    - Student consultants who performed system and software development at a deeper technical level.
    - Acted as team leaders for projects and initiatives.
- Student Information Processing Board (SIPB).
    - Supported access to computer facilities across MIT for students.
    - Originally responsible for services like dialup, www.mit.edu, discussion forum, printing, LaTeX support, and Linux-Athena.

Sustainability

- MIT, Digital, and IBM originally invested $100 million over an eight-year period (1983-1991) in hardware, maintenance, staff support, and software.
- Current staffing levels for IS&T to support Athena is one full-time equivalent (FTE).
- Student Information Processing Board (SIPB) also provides support for Athena to undergraduates.
- Availability of personal computers has drastically reduced the need for Athena workstation clusters.

In its heyday, Athena was a transformative force, and it spanned academic and enterprise IT activities in a synergistic partnership for several years. In a familiar pattern, as the academic elements of MIT transferred more resource operations to enterprise IT, support for Athena suffered. Today, Athena has become a largely stagnant resource that is not seriously addressed in any budgets.

TEACHING AND LEARNING WITH COMPUTATION AND TECHNOLOGY

As with the history of research computing, the transformative importance of digital technologies within education—in and beyond the classroom, in forms of teaching, learning, assessment, and with infrastructural support—has increased exponentially over the last 25 years. The MIT-wide launch of OpenCourseWare (OCW) in 2002 (i.e., sharing large amounts of course content online through a centralized site that is well-supported by individualized staff liaisons within every academic department) successfully modeled the new possibilities of the internet for transforming 21st-century education. OCW consciously involved all five

Schools and aimed to address both MIT's mission of service to the world and its own residential campus needs. OCW continues to honor those aims as it expands its educator services and YouTube presence.

In 2003, MIT President Charles Vest convened a task force focused exclusively on the Undergraduate Educational Commons to address those aspects unaddressed by the Life and Learning report of the Task Force on Student Life and Learning in 1996[3]. The 2003 task force, chaired by Dean Robert Silbey, studied all dimensions of the General Institute Requirements (GIRs) and set forth its recommendations in 2006[4]. In addition to changes in the Humanities, Arts, and Social Sciences (HASS) Requirements, new prioritization of international experiences, and a call to address the need for infrastructural improvements, the task force recommended a newly designed Science, Mathematics, and Engineering Requirement that included options for computation, design, and project-based learning.

The category of "computation and engineering subjects" focused on modes of thought and problem-solving tools associated with the computational modes of analysis and the engineering method. It explored the role of algorithmic and data abstractions as well as the use of imperative knowledge in designing computational solutions for theoretical and practical problems. Those would not simply be introductions to programming languages, but rather would provide a computational paradigm of reasoning and problem solving and were imagined to involve collaborations among departments and other academic units (pp. 47-8).

The 2006 recommendations were the most significant attempt to rethink the Science Core within the wider context of an MIT education since the early 1960s. The design challenge—adding new knowledge without unrealistic increases to an already constrained four-year required curriculum—proved contentious. The majority vote of the faculty recommended instituting its recommendations, but that did not constitute the super-majority needed for those particular curricular reforms to be endorsed as a unified package (the inclusion of choices among Science Core subjects proved to be particularly divisive).

[3] http://web.mit.edu/committees/sll/
[4] http://web.mit.edu/committees/edcommons/documents/task_force_report.html

An outward-facing transformation occurred in 2012 when the Provost and President Reif lead the establishment of MITx via edX (a collaboration with Harvard University) as a nonprofit MOOC delivery site. That initiative simultaneously built on and enriched online pedagogies within MIT's residential curricula. The reorganization of certain units to create a new Office of Digital Learning featured an infrastructural model that was run centrally but provided services and partnership to DLCs across the Institute. A major effort to consider the impact of technology and computing on teaching and learning at MIT was initiated by President Reif in the 2014: Task Force on the Future of MIT Education[5].

With the support of MIT's central administration, those major initiatives gave focus and inspiration to many local innovators in the use of computation within teaching and learning. Nonetheless, some educational projects did not fit within the parameters of OCW or MITx. Such projects remained either thriving independent local resources, languished, became obsolete, or disappeared because of a lack of infrastructural support. The SCoC provides a perfect opportunity to reassess the overall landscape of educational computing and AI infrastructure. Our working group views this moment as a time to consider supplementing MIT-wide initiatives such as OCW and MITx with greater efficiencies and networking of grassroots initiatives and essential learning tools.

The SCoC also provides an opportunity to unite such strides in infrastructure with a strong history of engaged faculty governance in advancing the curriculum and overseeing education. In 2016, a seven-person ad hoc Working Group on Algorithmic Reasoning and Computational Thinking was charged by then-Chair of the Faculty Krishna Rajagopal and Dean for Undergraduate Education Dennis Freeman to consider a set of questions related to that topic. The group, chaired by Prof. Eric Grimson, recommended that all undergraduates be required to take at least one subject offering in computation (Grimson Report[6], p. 1).

In addressing the question, "What is computational thinking?" our working group argued that computational thinking:

- Provides a distinct type of rigorous thought of important intellectual value.

[5] https://future.mit.edu
[6] https://facultygovernance.mit.edu/sites/default/files/reports/2017-01_computational_thinking_requirement_FINAL_CLEAN.pdf

- Requires and develops important modes of communication.
- Acknowledges the transformational impact of computation across many disciplines.
- Creates opportunities for MIT students and graduates.

Computational thinking involves more than the skill of computer programming or the ability to use computer tools. It includes fundamental modes of reasoning about rendering of physical or social systems in a manner that enables computational experiments to complement physical or social ones. Our working group additionally codified a set of required knowledge to enable computational thinking. Perhaps most relevant to this report, we established that computational thinking requires the use of a computer programming language as a framework within which to explore computational and algorithmic concepts. It also requires the ability to go from formulation to solution and to create or explore complex models of social, physical, or biological systems. These activities clearly must be enabled by significant computational infrastructure. Our discussions are ongoing, and we have not reached a clear consensus on how this thinking should be integrated within the curriculum across all departments.

The creation of the Office of the Vice Chancellor (OVC) in 2017—uniting oversight of undergraduate and graduate education—presents new opportunities. In fact, the charge by the Chancellor to the OVC parallels many aspects of previous initiatives in that MIT must continually evaluate its strengths and weaknesses with regard to the shifting global, technological, economic, and political landscape. In weighing the importance of MIT values and principles, faculty responding to a survey ranked hands-on experience second only to a commitment to excellence. Students ranked hands-on experience as most important, further emphasizing the need for renewed investment in computational infrastructure.

TERMINOLOGY

MEANING OF COMPUTING INFRASTRUCTURE

During our working group discussions, we found it necessary to define what we mean by computing infrastructure. It soon became obvious that no single definition would inspire universal agreement across the MIT campus. To some, computing infrastructure implies personal computing, emails, and office software. To others, it represents a variety of high performance computing options on campus, off campus, or on the cloud as well as the use of specialized software or the management/analysis of large databases. To help focus our report, we concentrated on the infrastructures needed to support research and education but excluded physical classroom, office infrastructure, and enterprise computing. Our discussions explored models related to:

- Computing resources (local clusters, shared central resources, cloud computing, etc.).
- Storage resources (local, central, cloud, etc.).
- Data management resources (policies, processes, multi-layered access, etc.).
- Common software tools for education and research and management strategies.
- Cyberinfrastructure for Sustained Scientific Innovation (e.g. centralized repositories, etc.).
- Support for scientific computing (support staff for setup and training, etc.).
- Interface between the SCoC infrastructure and existing MIT resources.

APPENDIX

RESEARCH COMPUTING SURVEY

384 respondents, 25% overall response rate.

(1) 65% had some computing needs beyond desktop/laptop.
50% < 1 million CPU or GPU hours/year
20% > 10 million CPU or GPU hours/year
6% > 100 million CPU or GPU hours/year

40% < 10TiB storage
13% > 100 TiB
4% > 1PiB

15% more expect to have needs beyond desktop/laptop in next 5 years
50% expect compute usage to stay or grow by less than a factor of 2 in next 5 years
13% (40 PIs) expect compute usage to grow by a factor of more than 5.
43% did not expect their storage needs to grow by more than 2 in next 5 years.
14% (43 PIs) expected their storage needs to grow by more than 5 in the next 5 years.

Several science and engineering departments expect x5 growth, but so do Economics, Architecture, DUSP, and Political Science.

(2) About 30% (83 PIs) were making some use of commercial cloud today.
30% of PIs were using national or other non-MIT partner resources.
70% of PIs were using some shared or individual local cluster.

(3) About 50% of respondents cited factors limiting the use of commercial cloud.
The most cited factor was cost. F&A surcharge was cited by 61 PIs.
23% of respondents were making some use of AWS cloud basic services.
22% of respondents were making some use of Google cloud basic services.

2% of respondents were using Azure.

4% of respondents were using cloud provider API services for advanced toolkits.

4% of respondents were using some other cloud provider.

(4) About 60% of respondents would find a low cost, large volume convenient storage solution valuable.

38% were interested in storage solutions for ensuring research reproducibility and for open data sharing.

16% (59 PIs) were interested in storage solutions that can be shown to be compliant with data use agreements (e.g., HIPPA terms, EAR, ITAR)

(5) Nearly all PIs would consider individually owned systems, collectively managed systems, or cloud systems if purchasing new hardware.

50 PIs would definitely not consider an individually owned system.

34 PIs would definitely not consider commercial cloud resources.

10 PIs would definitely not consider a collectively managed system.

(6) Open text responses suggested.

(a) More solid support for shared cluster services like the Engaging cluster and the CSAIL Openstack system.
(b) More solid support for training activities.
(c) More aggressive pursuit of cost-effective cloud services.
(d) Do not overlook the less traditional needs of humanities around video/audio/text processing, serving, and other specialized requirements. At least one commenter felt that there was a gap in thinking in this specific area between the desktop and large-scale facilities. The comment highlighted the importance of better measuring (and addressing) the needs of humanities research.
(e) General interest in growing base services available at no direct charge.

INFRASTRUCTURE WORKING GROUP SURVEY

The SCoC Infrastructure Working Group conducted a survey of MIT's DLCs on their computing infrastructure practices and needs. More than 20 DLCs responded to the survey, which surfaced the following major themes:

- Cloud versus on-site computational resources.
- Computational resource funding model.
- Infrastructure resource sharing.
- Access to computational resources for students.
- Growing need for machine learning services.
- Need for scientific computing support.
- Data access, management, and backup.
- Access to expert technical support.
- Interactive platform for computational services.

Cloud Versus On-Site Computational Resources

- All respondents cited use of both cloud and local computing.
- MIT has multiple centers for computing (i.e. high performance, genomic, machine learning, etc.).
- Cloud computing resources are increasingly used because of the breadth of their service portfolio.
- Cloud resources are currently seen as a more expensive option, but we will need to look at the total cost of ownership for running and supporting infrastructure in the cloud.
- Local clusters have suffered from reliability and recoverability issues at times. These issues were infrequent, but when they occur it severely impacted project work.

Computational Resource Funding Model

- More than 50% of the respondents would like to have access to lower cost storage and compute resources.

- DLCs stated they want computational resources to be treated as a commodity, like electricity.
- Partner with commercial cloud vendors to provide low-cost infrastructure and services. Similar to our past agreements with Athena and Digital/IBM.
- Aggressively negotiate terms with commercial vendors for computational resources.
- Some labs fund infrastructure through Consortia funds as centralized lab overhead.
- "If all your users are completely happy, you have probably overspent (by a lot)."

Infrastructure Resource Sharing

- Look at a way to pool computational resources across MIT. This is currently done, but on a smaller scale.
- Governance would be needed to avoid resource abuse and allow equitable access to shared resources.
- A researcher in one DLC was surprised that MIT did not have a central computational resource service.
- Allocation methodology of central pool resources to projects for priority access.

Access to Computational Resources for Students

- All respondents specified that providing computational resources for students is a must.
- MIT could leverage cloud vendors to provide resources to students at no cost.
- It is important to have access to digitized tools and opportunities for classroom education.

Growing Need for Machine Learning Services

- ~75% of respondents cited machine learning as a growing trend.
- More technical support is needed for developing and executing models.
- Need more GPU clusters to execute ML and AI applications.

Need for Scientific Computing Support

- A consulting group for scientific computing support.
- Support staff that would have research experience in specific domain areas.
- Support that would be able to develop tools and services that could be reused by many DLCs.
- PhD level support for computing.

Data Access, Management and Backup

- Services to provide long-term data storage.
- The ability to store, discover and retrieve data easily.
- Services to manage security requirements for data from sponsors (i.e. HIPAA, export control).
- Large data retrieval that is efficient and low-cost. Data egress costs can be an issue.
- Backup management for large datasets.
- Metadata management for proper data access.

Access to Expert Technical Support

- Currently, there is a severe shortage of technical support.
- Technical support for HPC, parallel, and cloud architecture should be available.
- Hire more engineering and support staff to support faculty development, students, and innovative pedagogy.
- Consulting model for DLCs that require additional support.
- Support must be sustainable to avoid an initial peak of service and then slow degradation.
- Can we reuse the Athena model, which was run by researchers?
- Excellent maintenance and support is critical and is as important as fast infrastructure.

Interactive Platform for Computational Services

- Creation of an interactive platform for computational resources that would lower the barrier of entry and provide easy access.
- Ability to develop and publish computational services (AI, ML, algorithms) for use.
- Platform should provide a consistent approach for managing data, resources and job execution.
- Platform should support parallel computational architectures.
- Modular platform solution that can use commercial offerings and MIT-developed components.
- Value should be for computational needs and computer science.

MIT ENTERPRISE RESOURCES AND SERVICES

MIT offers computing resources and services that support enterprise activities at the Institute: workplace services, connectivity, finance, human resources, student information management, facilities, office management, productivity applications, documentation, and others.

The table below highlights the major enterprise software and services made available to support the MIT community.

CATEGORY	SOFTWARE / SERVICE
Administrative	SAP HANA (Finance) SAP HANA (Human Resources) SAP HANA (Facilities) Coupa Concur Kuali Coeus ServiceNow Ellucian Advance eBuilder (Facilities)
Education	MITSIS Stellar LMOD

	Banner Admissions
Workplace	Dropbox Microsoft Office Microsoft Exchange (email and calendar) Citrix Code 42 CrashPlan Data Warehouse IBM Cognos Tableau Lynda.com WebEx DocuSign
Community	Oracle Microsoft SQL Server VMware Qualtrics GitHub Quickbase FileMaker Salesforce Drupal Tivoli Storage Manager MIT Mobile App Wikis Xfinity on Campus
Network / Security	MITNet Wireless Network Duo Security Sophos Anti-Virus CrowdStrike Lastpass Spirion (formerly Identity Finder) Cisco UCC Kerberos Touchstone Authentication OpenID

EXAMPLES OF CURRENT STATE FROM DLC SURVEYS

Civil and Environmental Engineering Survey

- "One issue is the clock-tick of research and administration. To achieve this the organizational structure needs to be very different. For example, OSP operates at a clock-tick of months to put contracts in place. Similarly with TLO, the lag time is weeks not days or even hours. The SAP system means that we employ layers of financial 'interpreters' who insert themselves between the PI and the data."

Electrical Engineering Survey

- "Can we create some uniform standards for how data is stored and shared that would make it easier to try out projects? We need to move the ramp-up time down from ~months to ~1 week. Low barrier to entry is critical."

Physics Survey

- "Considering the large need for data not only in physics experiments and the long term storage requirements it is desirable that MIT establishes a Data Management Plan and Infrastructure."

MIT Brain and Cognitive Sciences Survey

- "Sur lab generates 500gb a day of data; uploading and downloading takes days; terribly inefficient; accessing data takes days and weeks; we need faster pipes."
- "Jazayeri Neuroscience Lab is now moving to recording thousands units simultaneously; big worry down the line because we don't know how to handle the quantity of data. As size of data becomes large; [Could MIT] think about model where data is available nationally, giving people access to data instead of saving data locally?"
- "Data back-up is a problem; when it goes down, takes forever to restore."

- "Making experimental data accessible is becoming more widespread. However, managing access and (equally important) tracking subsequent use (or mis-use) is challenging. It would be good if the SCoC could provide support, standardized access protocols, etc."

CSAIL's Discussion of Key Issues and Suggestions for the College of Computing

- "Security constraints on data (such as HIPAA compliance) is a major issue, as it makes sharing data challenging. Identifying or developing computing systems that are able to provide pre-defined ways of meeting such requirements could be very helpful."
- "For students taking courses who want to do a project that combines large data sets, it can take too long for them to set up an account. They run out of time before the course is over. It would be very helpful campus-wide to have the ability to mount any of your Deep Learning platforms or some sort of huge data storage, etc., in a kind of seamless way in Athena or whatever other kind of campus-wide computing platform there is without having to go through the hassle of accounts."
- "Companies provide data for research but require constraints, and medical data has limitations too, etc. Being able to meet those constraints but also share data is a challenge. It would be helpful for the computing system to have pre-defined ways of meeting whatever security requirements are needed (such as being HIPAA compliant) by ticking a box when setting up."

Civil and Environmental Engineering Survey

- "Computational biology... amount of data being generated per dollar keeps increasing. Need cool ways to leverage data for insights."
- [Pain points include] "compiling and linking code to clusters; Software licensing and installation; storing data (from lab experiments and computer simulations)."

SHASS Survey

- "There is real interest among our faculty and graduate students to learn more about and have reliable access to data visualization tools, the collection and use of big data sets, visual anthropology, etc. Access to such technologies and methods is increasingly important in graduate admissions and faculty recruitment. Peer institutions are ahead of us on this."

MIT Kavli Institute Survey

- "The TESS spacecraft project currently has well over a PB of disk for storing and backing up data; this will continuously grow. For full exploitation of LIGO data the ability to store ~full-bandwidth data (O(1)PB per year) will be needed. The ideal case would be that of a storage service, where backups are centrally sourced and managed. Long-term storage using tape systems is highly desirable. One important need is for verifiable EAR/ITAR/PII compliant storage. We urge that this be supported alongside other compliance tools (e.g. HIPAA) in any data storage solution, as imposed by sponsoring agencies and/or state or federal law."

Political Science Survey

- "I've been paying for cloud storage for large administrative datasets (about 600 GB) that allow for easy and fast access. Would be great to have a local solution, both for the storage and the database access/maintenance."
- "Our primary need, beyond what we have now, is access to more massive storage. In the process of capturing election data, we are already having to shift around data in order to accommodate space limitations, either on personal machines or MIT-based resources. At the moment, we do not see computational power as being a significant constraint. The ideal scenario, though, would be to have free or cheap access to more working data storage space."

Physics Survey

- "When the data is considered very large, like for the LHC experiments (hundreds of PB), there is a detailed plan for storage and data management in place which is

supported through grants with DOE and NSF. For other experiments with smaller data volumes local support is more difficult and the research groups could benefit from well managed local storage and data management concepts … Right now, in the physics department common data management software is mostly absent. The various entities in the department like headquarters, the fiscal offices in LNS and MKI have some software to manage their administrative data but it is implemented using mostly standard Microsoft based tools or more specialized software provided by MIT."

Mechanical Engineering Survey

- "Cyberinfrastructure for Sustained Scientific Innovation (e.g. centralized repositories, …), for example expanding the central digital archive, a repository at the Institute level, allowing digital works, reports, audio, video, data, and more to be shared and securely stored."

SHASS Survey

- Pain points include "data storage and durable archival hosting of videos (especially after the demise of Tech TV) and project websites."
- "In the humanities, just as in engineering and the sciences, we have a need for Institute-supported, long-term data conservation … we have [migrated our data to freely-available international repositories in France and Canada] … It is difficult to explain to our international colleagues why MIT has been unable to provide a solution to this problem."

Libraries Survey

- "Although mature methodologies and services exist to support data procurement, management, storage, and archiving (as well as small-scale hybrid infrastructure for long-term archiving and preservation of data), providing data services and support at the scale anticipated by the SCoC will require partnerships across the institute between the libraries, IS&T, and research labs and centers across MIT."

Political Science Survey

- "For my local government transparency project (this is a project on local US governments; how many make items like their budgets available online, etc.), I'd like to start archiving some subset of the websites I scrape which would require large databases. Up to several terabytes I imagine."

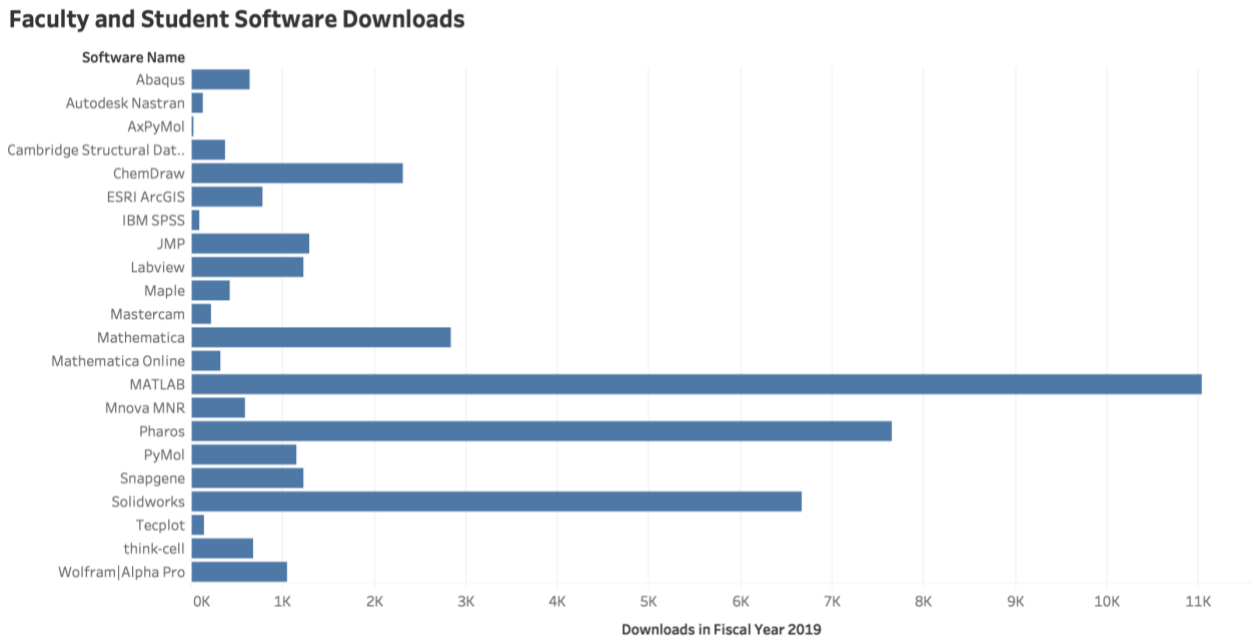
Physics Survey

- "Data storage at [exabyte scale] including hot and cold storage certainly requires a heavily integrated architecture and software stack including state-of-the-art tape archival capabilities."

CURRENT SOFTWARE PACKAGES

The current software packages are categorized as follows:

- Bioinformatics (15+)
- Finite Element Analysis (10+)
- Graphics - CAD / Video / Drawing (40+)
- Geographic Information Systems (2)
- Molecular Modeling and Dynamics (25+)
- Numerical / Math Software (30+)
- Programming (90+)
- Simulations (25+)
- Other (50+)



In addition to open source software, the following commercial software was highlighted in the MIT PI research computing survey

- MATLAB, 45% of research computing survey respondents
- Compiler and debugging suites from Intel, NVidia and others, 27% of research computing survey respondents
- Mathematica, 25% of research computing survey respondents
- Globus and other data management and data transfer automation tools, 10% of research computing survey respondents
- Solidworks
- ChemDraw
- Dropbox
- Labview
- Ansys
- COMSOL
- Stata
- ArcGIS
- Adobe Creative Suite (Photoshop, Illustrator, Acrobat)
- Microsoft Office

A full list of software identified by the survey is listed in ***Appendix A. MIT PI Research Computing Survey - Software in Use***.

PEER INSTITUTION ANALYSIS

- Stanford
  https://srcc.stanford.edu/
  - 20 FTEs in research computing https://srcc.stanford.edu/about/people
  - Overall 20 petabytes of data, 2,000 HPC servers, 30,000 CPU cores, 2,040 GPUs, 13,000 user accounts
  - Research HPC has 1,325 compute nodes, 24,096 CPU cores, 1,195 GPUs and 1,590 TFlops of computing power https://srcc.stanford.edu/sherlock-high-performance-computing-cluster
  - Specialized cluster for secure data: http://med.stanford.edu/nero/why-nero.html
- Berkeley
  http://research-it.berkeley.edu/programs/berkeley-research-computing
  - 24 FTEs http://research-it.berkeley.edu/about
  - 470-node, 11,620 processor-core Linux cluster rated at nearly 450 peak teraFLOPS http://research-it.berkeley.edu/services/high-performance-computing
  - Has on-campus support for AWS/Azure/GCP
  - Provides consulting http://research-it.berkeley.edu/programs/consulting
  - Specialized service for managing data: http://researchdata.berkeley.edu/services
- Carnegie Mellon University (CMU)
  https://computing.cs.cmu.edu/research/index.html
  - Unclear how large the campus team is
  - No specifics for on-campus clusters
  - Pittsburgh Supercomputing Centre has 48 FTEs and 6 executives
  - PSC provides consulting https://www.psc.edu/resources/consulting
  - Unclear size of resources
- Harvard
  https://www.rc.fas.harvard.edu/
  - 24 FTEs https://www.rc.fas.harvard.edu/about/people/
  - Odyssey3 has 78,000+ cores, 35 PB of storage, 2000+ Nodes, 250+ TB of memory, 1M+ CUDA cores
  - Does not appear to offer centralized consulting, appears more to be within individual DLCs
  - Does not appear to have substantial data management services
- Princeton
  https://researchcomputing.princeton.edu/

    - 29 FTEs, 7 executives https://researchcomputing.princeton.edu/people
    - 30,000 total cores, 6PB of storage and over 2,000 TFLOPS https://researchcomputing.princeton.edu/systems-and-services
    - Provides consulting services https://researchcomputing.princeton.edu/systems-and-services
    - Does not appear to have substantial data management services
- Columbia
    - Could not get data on staffing
    - Four clusters of various sizes https://cuit.columbia.edu/shared-research-computing-facility
    - Has on-campus support for AWS https://cuit.columbia.edu/aws with a negotiated agreement
    - Offers data management services https://cuit.columbia.edu/sde
- University of Washington https://itconnect.uw.edu/research/
    - 6 FTEs https://s3-us-west-2.amazonaws.com/uw-s3-cdn/wp-content/uploads/sites/70/2019/04/15115506/UW-IT-Org-Charts-for-Presentations-04.15.19.pdf
    - 10,784 cores, 828 nodes
    - Has on-campus support for AWS and Azure with umbrella cloud agreements https://itconnect.uw.edu/research/cloud-computing-for-research/getting-started/
    - Unclear how much storage is available but is primarily an archive service rather than a data management service https://itconnect.uw.edu/service/shared-central-file-system-for-research-archives-lolo-archive/
    - Provides consulting services https://itconnect.uw.edu/research/hpc/research-computing-services/#outreach
    - UW provides the ability to get F&A waivers for computing and storage: https://itconnect.uw.edu/research/waiver/

ADDITIONAL INPUT FROM TEACHING AND LEARNING STAKEHOLDER GROUPS

What Works Well for You Right Now

- MGHPCC model has been working very well from a PI/department perspective. Purchased nodes are given a high priority to the buyer but access is granted to everyone when not in use. Basic technical support is provided to migrate codes,

data, and users to the platform. There are non-recurring costs associated with hosting nodes at MGHPCC.

- There is a heavy reliance at MIT on government-backed high performance computing centers (NSERC, ALCF, OLCF). These resources provide very good technical support but access can sometimes be limited or through competitive processes.
- HPC computing has advanced the physical understanding in the basic science and engineering fields and is now a foundational tool for further scientific advancements.

What Are Some of Your Lessons Learned

- Departments and units are often too small to support the dedicated technical staff. We saw a reliance on generous researchers to provide necessary support in addition to their workload.
- Centralized model helps provide a secure environment that is easier to maintain and keep up-to-date.
- Training through IAP on how to use current resources has been proven very effective, but ad hoc department-led efforts are not always sustainable.
- Migration to newer computational paradigms is an active research field supported by the government (e.g. Exascale computing project).
- Looking forward, many people expect increasing integration between traditional large-scale HPC simulation and tools from AI and machine learning. This can be as straightforward as unsupervised learning to identify patterns and structure, or as complex as trying to use learning approaches to develop new methods for solving complex equations sets. In all cases there are implications for infrastructure and also for cross-domain knowledge and collaboration.

Pain Points

- Machine is never large enough to accommodate research growth and higher fidelity models. No matter how large the resources you provide, they will never be large enough. Research will expand to fill the space.

- Queues are always filled with single processor jobs.
- There is no sustainability plan to replace the purchased nodes once it becomes obsolete. This creates a space problem that prevents others from buying or expanding, it encourages keeping obsolete nodes running past their optimal lifetime. Space constraint may push PIs to start bringing resources back to campus in their own labs.
- No financial mechanism to accumulate computing funds for renewing infrastructure. No incentive to get rid of old nodes.
- Job turnaround time is quite slow on centralized systems which pushes PIs to buy their own resources even when not always in use.
- Skills training for large scale computing could be more widely available.

COMPUTING INFRASTRUCTURE WORKING GROUP MEMBERS

Benoit Forget (Co-Chair)
Associate Professor, Department of Nuclear Science and Engineering

Nicholas Roy (Co-Chair)
Professor, Department of Aeronautics and Astronautics

John (Jack) Costanza
IT/IS Manager, Computer Science and Artificial Intelligence Laboratory

Martin Culpepper
Professor, Department of Mechanical Engineering

James Damewood
Graduate student, Department of Material Science and Engineering

Lisa George
Senior Director, IT, Office Development Systems, Office of Resource Development

Diana E. Henderson
Professor, Literature Section

Christopher Hill
Principal Research Scientist, Department of Earth, Atmospheric and Planetary Sciences

Marc Jones
Assistant Dean for Finance and Administration, School of Humanities, Arts, and Social Sciences

Frans Kaashoek
Charles A. Piper (1935) Professor of Computer Science and Engineering, Department of Electrical Engineering and Computer Science

Jeremy Kepner
Laboratory Fellow, Lincoln Laboratory

Retsef Levi
J. Spencer Standish (1945) Professor of Management and Professor of Operations Management, Sloan School of Management

Yousef Marzouk
Associate Professor, Department of Aeronautics and Astronautics; Member, Institute for Data, Systems, and Society

Josh McDermott
Frederick A. (1971) and Carole J. Middleton Career Development Assistant Professor of Cognitive Science, Department of Brain and Cognitive Sciences

Michael P. Rutter
Senior Advisor for Communications, Office of the Vice Chancellor for Undergraduate and Graduation Education

Mark Silis
Associate Vice President, Information Systems and Technology

Krystyn Van Vliet
Associate Provost; Michael (1949) and Sonja Koerner Professor, Departments of Materials Science and Engineering and Biological Engineering

John Williams
Professor, Department of Civil and Environmental Engineering

Sarah Williams
Homer A. Burnell Career Development Associate Professor of Information Technologies and Urban Planning, Department of Urban Studies and Planning; Member, Institute for Data, Systems, and Society

Boleslaw Wyslouch
Professor, Department of Physics

Heather Yager
Associate Director for Technology, MIT Libraries