The Discovery Cloud: Accelerating and Democratizing Research on a Global Scale

Ian Foster*†‡, Kyle Chard*, Steven Tuecke*
*Computation Institute, Argonne and UChicago, Chicago, IL 60637, USA
†Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, USA
‡Department of Computer Science, The University of Chicago, Chicago, IL 60637, USA
Email: {chard, foster, tuecke}@uchicago.edu

***Abstract*—Modern science and engineering require increasingly sophisticated information technology (IT) for data analysis, simulation, and related tasks. Yet the small to medium laboratories (SMLs) in which the majority of research advances occur increasingly lack the human and financial capital needed to acquire and operate such IT. New methods are needed to provide all researchers with access to state-of-the-art scientific capabilities, regardless of their location and budget. Industry has demonstrated the value of cloud-hosted software- and platform-as-a-service approaches; small businesses that outsource their IT to third-party providers slash costs and accelerate innovation. However, few business cloud services are transferable to science. We thus propose the Discovery Cloud, an ecosystem of new, community-produced services to which SMLs can outsource common activities, from data management and analysis to collaboration and experiment automation. We explain the need for a Discovery Platform to streamline the creation and operation of new and interoperable services, and a Discovery Exchange to facilitate the use and sustainability of Discovery Cloud services. We report on our experiences building early elements of the Discovery Platform in the form of Globus services, and on the experiences of those who have applied those services in innovative applications.**

1. Introduction

Research faces a crisis due to rapidly growing data volumes, use of computational methods, collaborative research, and methodological complexity—factors that both individually and collectively threaten the viability of the small and medium laboratories (SMLs) within which most research occurs and most discoveries are made. SMLs typically lack the budgets and staff needed to develop, maintain, and apply the specialized IT infrastructures, methods, and tools that are increasingly required for research success. The resulting growing gap between IT needs and capabilities makes researchers less competitive and impedes technology transfer and innovation. It also disenfranchises growing numbers of researchers worldwide, at a time when the need for innovation has never been greater.

We thus propose the *Discovery Cloud*, a new cyberinfrastructure ecosystem that will deliver needed state-of-the-art IT capabilities to SMLs in a manner that is easily usable, scalable, and economically sustainable. This ecosystem will (a) address directly the data lifecycle and computational challenges faced by SML researchers; (b) allow rapid adoption within SMLs, without requiring substantial technical expertise or local infrastructure; and (c) adapt easily to varied disciplinary needs and rapidly evolving science requirements. In so doing, it will enable state-of-the-art discovery and education within those labs that today struggle with IT challenges, while also accelerating work within even the best-resourced labs by slashing time spent on mundane and routine activities. It will thus allow SMLs to compete and indeed prosper in a world of increasingly large, noisy, and complex datasets and ever more sophisticated computation.

In proposing the Discovery Cloud, we leverage recent advances in commercial IT, notably cloud-hosted software as a service (SaaS) [1]–[3] and platform as a service (PaaS) [4]. These new methods are widely used to streamline and accelerate operations within small and medium businesses (SMBs), allowing them to outsource to third-party providers many (often the entirety) of the business processes that used to be a major source of overhead and inefficiency. These capabilities allow the Discovery Cloud to focus on developing, adapting, and integrating higher-level services, including those that were originally developed for enterprise or big science collaborations, and delivering the resulting tools through a framework designed specifically to meet SML requirements. The resulting Discovery Cloud will both provide direct support to SMLs and catalyze the creation of a broader ecosystem of research services to support a wide range of SML needs.

While our vision for the Discovery Cloud is expansive, it is firmly rooted in practical experience. Over the past five years we have developed and operated Globus [5]—a collection of cloud-hosted services designed specifically to address SML data management challenges. The results of this somewhat radical experiment have been extremely positive: Globus services have been incorporated into the research workflows of many thousands of people and projects worldwide. The rapid growth of usage combined with overwhelmingly positive feedback, has highlighted the value of service-based delivery of scientific capabilities.

The Discovery Cloud that we propose here will build on this experience, leveraging proven Globus services (e.g., authentication, identity management, and data transfer), while supporting the development and integration of new capa-

978-1-5090-1961-8/16 $31.00 © 2016 IEEE
DOI 10.1109/IC2E.2016.46

bilities, providing a platform to simplify the creation of new services while reducing development and operations costs, and creating new mechanisms that directly address the inherent sustainability challenges of research software.

The rest of this paper is as follows. In §2 and §3 we introduce SMLs and characterize their cyberinfrastructure needs. In §4 we review existing cyberinfrastructure approaches, outline why those approaches are unsuitable for science, and examine what we can learn from commercial software. We describe our vision for the Discovery Cloud in §5 and present an architecture for delivering this vision in §6–8. In §9 we discuss sustainability challenges for scientific software. We summarize in §10.

2. SMLs: The powerhouse of science

We first define and review what we know about the target community for Discovery Cloud: the small and medium laboratories (SMLs) in which most research is performed.

We can gain insight into research lab sizes by studying the sizes of federal grants. A 2007 analysis of National Science Foundation awards showed that when ordered by size, small and medium grants (those less than $1M in total) account for 80% of total dollars and 98% of all awards [6]. A subsequent analysis by Weber [7] found similar results in later years. Similarly, analysis of National Institutes of Health awards reveals an average award amount of $431,177 in 2014 [8]. We see a power law distribution in which a few awards are extremely large but the majority are small.

Presumably some researchers will have multiple grants and some will have funding from other sources. On the other hand, many researchers have no federal funding, even at research universities, and many scientists and engineers work in other contexts. Indeed, the National Science Board estimated four to eight million science and engineering positions nationally in 2003 [9].

While these data do not provide definitive numbers on lab budgets, the overall impression, also supported by anecdotal evidence, is that most research labs operate on limited budgets (low $10,000s to $100,000s per year). We conclude that a substantial portion of research in the life sciences, physical sciences, and social sciences occurs in the tail of this power law distribution of resources. We further believe that most output comes from smaller producers. This view is supported by Fortin et al., who found that research productivity was not improved by concentrating funding into fewer, larger labs [10].

We use the term *small and medium laboratories* (SMLs) to refer to these powerhouses of research. The term is chosen by analogy to the term small and medium business (SMB: in Europe, small and medium enterprise, SME). Conventionally, a small business employs up to 50 people and a medium business up to 250. No conventional definition exists for SMLs, but we suggest that any group with a research budget of up to $1M/year is in scope. That number clearly encompasses the majority of federally-supported research. This budget level can support a single PI and a few students and postdocs, or in some cases several PIs and some research staff. Yet it is of a scale that makes supporting a dedicated IT team challenging or impossible.

3. Cyberinfrastructure needs of SMLs

We see much evidence that in an era of massive data and computation, competitiveness demands significant IT. We note important areas that highlight SML needs.

3.1. The big data challenge

Data challenges that used to be peculiar to big science fields such as high energy physics are now commonplace across disciplines as, for example, the cost of genome sequencing drops precipitously, data rates from synchrotron light sources, microscopes, telescopes, urban sensors, and other devices increase; and high-throughput methods create large databases of experimental and derived data. This so-called fourth paradigm [11] of scientific discovery (experiment, theory, and simulation are the first three) requires new tools and expertise.

Meanwhile, we observe that the hardware, software, and expertise available in typical small labs cannot cope with data at these levels of scale and complexity. There is much evidence to suggest that time spent handling data is already excessive. For example, a survey of 200 researchers reported spending at least 40% of their research time working with data, about a third of that time performing tasks unrelated to their field of study [12]. Further, over 60% of respondents reported that a lack of software tools sometimes prevents pursuit of particular research questions.

Similarly, it has been suggested that researchers and others working in data-intensive fields spend 80% or more of of their time wrangling data [13], [14]. As data volumes continue to grow, the likely outcome is that SMLs become incapable of engaging in leading-edge science.

3.2. Complex, distributed resource environments

Given the highly decentralized nature of the research community itself, and the fact that SMLs rarely have the resources necessary to acquire and operate sufficiently powerful research infrastructures themselves, they frequently end up relying on external resources: department clusters, campus research computing centers, national supercomputer centers, commercial cloud services, etc., for computing; remote genome sequencing facilities, MRI machines, and other research facilities; and local and remote systems for data storage and backup. The resulting complexity puts an additional burden on research tools and services: not only do they need to support the methodology required for the science itself, but they also must be able to operate across a distributed, dynamic environment.

3.3. Evidence from research services

Experience with Globus provides insights regarding demand for research data management services. In five years of operating the Globus service, more than 30,000 users have

registered, and more than 1,500 distinct users access the service each month. Over 20 billion files have been processed and more than 100PB of data transferred. As shown in Figure 1, usage continues to grow, driven by a mix of SML users, research institutions working to support their SMLs, and larger projects seeking to streamline their operations by automating operations. Other research services, such as iPlant [15] and nanoHUB [16], have similar stories to tell. This level of adoption suggests that there is a substantial and growing need for automating data management and that users see great value in using high-quality services.

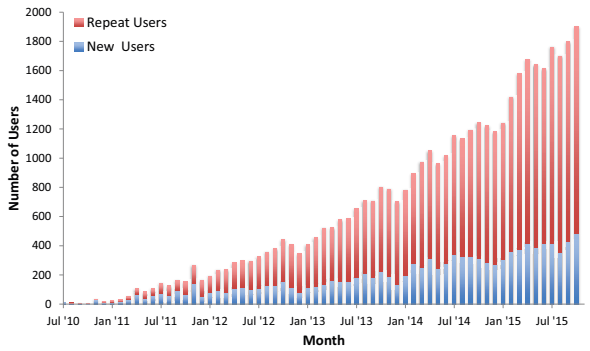

Figure 1. Globus research service usage, showing the number of active users per month (first-time and repeat users).

4. Current approaches

We have presented evidence that SMLs lack the human, financial, and cyberinfrastructure resources needed to exploit new opportunities associated with the use of advanced computation and big data. We now turn to the question of how these needs can be met. We review today's major research infrastructures and discuss how SML needs are not met by conventional software distribution methods and commercial software. We then identify the lessons to be learned from industry's adoption of SaaS and how it has improved the environment for small and medium-sized businesses (SMBs).

4.1. Today's cyberinfrastructure

The term *cyberinfrastructure* encompasses national research computing facilities such as XSEDE [17] and Open Science Grid (OSG) [18]; data infrastructures such as that provided by DataONE [19]; network infrastructure provided by the likes of Internet2, ESnet, and campuses; domain-specific infrastructures such as the Long Term Ecological Research (LTER) network [20], high energy physics experiments, astronomy projects, and the like; elements of experimental facilities, such as those operated by the Department of Energy (DOE); and computing and storage resources provided by campus research computing and IT organizations. Increasingly, commercial cloud services also form an important element of the national cyberinfrastructure [21].

These cyberinfrastructure components each provide important services to various elements of the research community. However, they are far from providing a complete answer to SML needs. Science-driven cyberinfrastructure investments have tended either to provide a unique capability (e.g., XSEDE: supercomputers) or to meet the specialized needs of one community (e.g., OSG: high energy physics data processing). This narrow focus means that each system supports relatively few users. For example, XSEDE supported 2,679 individuals in 2015, plus another 2,874 via various science gateways.

Today's cyberinfrastructure also displays a remarkable lack of commonality in technology. Individual systems tend to be both discipline-specific and implemented as vertically integrated, non-interoperable stacks. This diversity introduces major additional challenges for SMLs: the heterogeneity and frequent complexity (due in part to one-off, under-resourced solutions) increases the complexity of the cyberinfrastructure environment.

A few cyberinfrastructure services do aspire to provide general solutions to many SMLs: for example, Globus, Galaxy [22], nanoHUB [16], and SQLShare [23]. In general, these services leverage the SaaS methods that we advocate to reach many more researchers.

4.1.1. Conventional methods. One approach to meeting unaddressed SML needs is to develop new software from scratch. Another is to fund research programmers within existing research groups to acquire, support, improve, and maintain the software required for each research group's work. To some degree, this is the approach taken within communities such as high energy physics, in which hundreds of software developers worldwide work collaboratively to develop complex software stacks.

Neither of these approaches is feasible for the majority of scientific software development, due to a reliance on overly expensive and non-sustainable software delivery models and the mismatch between conventional software development models and the often ill-defined and dynamic needs of science communities. The approach often followed by big science projects—forming a large team to build a custom project-specific solution to their data and computation challenges—is impractical for SMLs. Nor can SMLs simply adopt software developed by big science projects, as that software is typically constructed for use in specialized environments and within large teams. It is *enterprise software* (expensive, heavyweight). SMLs need the scientific equivalent of consumer or SMB software (low-cost, lightweight).

4.1.2. Today's commercial software. Another possible solution to SML IT challenges is for SMLs to adopt commercial solutions. With the rapid progress being made in commercial IT services, one might think that scientists could meet their needs using off-the-shelf products. And indeed, researchers do make extensive use of services such as MatLab, Google Docs, DropBox, Box, Amazon Web Services (AWS), BlueJeans conferencing, and GitHub. But for various reasons, important elements of the research lifecycle are

not supported well by commercial solutions. Furthermore, the following factors lead us to believe that these elements are unlikely ever to be supported by commercial solutions.

Cultural and methodological: Science is heavy on deep subject matter expertise and a culture of evidence-based decision making and mathematical rigor. But researchers lack funding, expertise, and appetite for significant IT investment. Business, in contrast, understands deeply the importance of IT investment, but has only recently begun to appreciate data-driven decision-making and careful experiment design. Yet science has advantages business does not: In particular, an explicit mandate to share ideas and deliverables. We believe that this mandate can allow a platform like Discovery Cloud to succeed where it would not in industry.

Fragmented, long-tail market: Commercial products are designed for specific markets: groups of customers with similar needs that can be served by a single product. For a market to be attractive to a commercial provider, it must have potential for considerable revenue. The science market for software does not reach this threshold. It is relatively small in total numbers and financial resources. Furthermore, it is fragmented, involving many modest-sized communities of similar interests (e.g., fields of study), each with specialized instruments, datasets, methods, and tools. A consequence of this diversity is that no single organization or product can serve all software needs of the market.

Highly distributed and collaborative trust models: The global research community spans thousands of independent employing institutions, hundreds of thousands of projects, and hundreds of funding agencies, with no central authority or hierarchy. Any combination of scientists may choose to work together, but only if they can create and enforce limited trust relationships [24], [25] supported by the tools they use. There is no central authority for establishing and maintaining the individual identities, credentials, groups, and other attributes required to manage these trust relationships. Services must be able to recognize identities certified by employing organizations, make reasonable trust decisions based on user-driven policies and consents, and enforce usage and access policies in the absence of a central authority. (Business environments may also involve cross-organizational cooperation, but in general business trust models are considerably simpler.)

Distributed and heterogeneous infrastructure: The infrastructure available to scientists includes national and international facilities, campus-based research computing centers, project-based facilities, lab resources, scientific instruments, and commercial cloud providers. Software for science communities must operate across various cross-sections of this infrastructure. (Business environments may also be distributed and heterogeneous, but rarely SMB environments.) We see this characteristic of science as fundamental to its distributed nature: data will always be produced in many locations, and expertise will always be found in unpredictable places.

High-speed network connectivity: Regional, national, and global research networks often provide orders-of-magnitude higher bandwidth than that available on business and consumer networks. This situation allows for fundamentally different approaches than can be contemplated by most commercial data services, which are generally designed for slower/lower-bandwidth networks, and often do not scale well on high-bandwidth networks.

4.1.3. The current research service marketplace. Many services that researchers require are already available in one form or another. Industry has already delivered numerous services for supporting collaborative activities, such as email list managers and hosts, website and wiki hosting, and shared folder and file support (scaled for use with documents, not big data), as provided by companies such as Google, Dropbox, Microsoft, and Facebook.

Another category of services that SMLs can quickly leverage from the general business market is those that allow users to quickly build and scale up a computing system. Amazon, Google, and Microsoft all offer cloud computing systems that allow users with little or no personal computing capacity or expertise to construct medium- and large-scale computing systems.

There are also an increasing set of research services, such as Globus, that play an important role for SMLs. Globus provides foundational services for authentication, authorization, identity management [26], groups, data management [27], and data publication [28].

The ability to execute repetitive or complex series of tasks automatically is increasingly critical to scientific activity. Services, such as Galaxy, provide workflow systems for automating these tasks. Many people use Galaxy in conjunction with Globus services for genomic research and other purposes [29].

Many other services are currently available—and used to greater or lesser degree—by SML researchers. These services meet important, general needs in SML research, but are not yet integrated with the rest of the SML research environment and thus are harder to use than they should be. Examples of such services include high-performance computing and high-throughput computing services, long-term data archival services, and big data platforms (e.g., Spark, Hadoop, Flink, GraphLab/Dato, Myria, Asterix, Splunk).

One notable SML need that has not, to our knowledge, been addressed sufficiently by any product is the digitization and/or conversion data from the pre-digital era: for example, historical ecological observational data, retrospective records from experiments conducted in SMLs, and data encoded in non-standard schema.

These latter two sets of services—those not yet available to SMLs and those available but hard to use—point to the need for a common platform for SML services. An appropriately designed platform can both simplify the process of developing new services (e.g., digitization services) and streamline the integration of existing services with each other and with the research environment, improving ease-of-use and accessibility.

4.2. Lessons from commercial software

While commercial software does not provide a complete solution to SML IT needs, it teaches important lessons concerning SaaS as a means of software delivery. SMBs need IT capabilities for basic operations: capabilities that are essential to their functioning, but are not typically a means of differentiating themselves or their products from competitors. Simplicity and cost are thus major drivers for implementation choices, and over the past decade, we have seen increased outsourcing of billing, payroll, web presence, email, customer relationship management (CRM), and other functions to third-party providers. This outsourcing then allows IT teams in SMBs to focus on core business issues. This lesson serves equally well in research laboratories: the laboratory is best served when its in-house staff (researchers and assistants) are empowered to focus on research issues without being distracted by basic IT problems.

Commercial SaaS has also transformed consumer engagement with IT, via products such as Google Mail, Google Docs, Flickr, Netflix, and Kayak. In these contexts, SaaS does more than supplant existing IT: it also allows for the emergence of new classes of more specialized services that build on top of basic services. For example, Tripit collects, integrates, transforms, and monitors information about an individual's travel plans. Using published interfaces provided by other SaaS providers, it integrates relevant information about your destination (e.g., weather), monitors flights for delays, and tells you if friends will be in the area. The resulting data (your travel plan) can then be examined via an intuitive modern interface, or interfaced to other devices such as text messages, dedicated mobile application platforms, and calendar services. Tripit's intuitive interface, deep comprehension of the problem domain, social networking features, and integration with personal information sources all contribute to its success, but the business model itself is only feasible within today's rich SaaS ecosystem.

An important consequence of SaaS in business is that it has leveled the playing field for smaller players. For example, before salesforce.com, only large enterprises could afford CRM systems, which gave those large enterprises a competitive advantage. SaaS CRM leveled that playing field, allowing SMBs to be more competitive. We believe this lesson also translates to science.

5. Imagining a SaaS-based SML ecosystem

As we noted above, commercial SaaS is used in research when SML requirements overlap with those of SMBs. However, no comparable SML-focused services exist today that would address unique characteristics of SML work processes, such as ill-structured data, exploratory and iterative computational analysis, ad hoc sharing, and the need for low-overhead but effective security. Yet such services can easily be imagined.

5.1. Life in a SaaS-based SML ecosystem

Consider this example: *A research data integration hub allows data in different formats to be submitted manually, or alternatively harvested by monitoring storage locations, email lists, and Wikis registered by an individual, research lab, or community. As data are identified, the hub analyzes them, extracts metadata, annotates files, and constructs data models. It applies tools to identify patterns in data and suggest unexpected connections across datasets. Users can subscribe to datasets of interest. Similar datasets can be flagged for comparison or aggregation into larger collections. RSS feeds alert users to the arrival of data similar to their recent deposits or search results.*

The potential of SaaS for science has not gone unnoticed, as evidenced by development of research portals and hubs and by the increased use by researchers of services such as Google docs, AWS, and Globus. SMLs appear to be attracted to SaaS for two reasons. The first is cost: for reasons that we discuss below, it is far cheaper to obtain many functions from SaaS services than to implement and operate them locally [3]. Second, SaaS can provide services that an SML cannot easily install, configure, and operate. For example, SMLs often over-rely on spreadsheets for data capture and analysis, rather than taking on the initially more complex, but ultimately far more effective, method of setting up and maintaining a relational database server.

5.2. Discovery Cloud components

We believe that three main elements are needed to enable a cyberinfrastructure ecosystem for SMLs.

First, we need *a broad suite of research services*: services that are useful and, indeed, necessary for research to be accomplished in a modern, collaborative, and data-intensive fashion. These services described in §4.1.3 provide a wide range of capabilities and come from a wide range of providers: commercial and non-profit, scientific and business-oriented. Some of these services are already provided in a manner amenable to SMLs, but many require new channels for awareness and accessibility within the SML environment.

In order to facilitate an ecosystem of independent research services we require a platform on which services can be developed and integrated. We call this second element the *Discovery Platform*. The Discovery Platform provides common capabilities that can be leveraged by other services, for example, common identities and groups, a flexible authentication model, analysis capabilities, and the ability to reference, access, and move distributed data. Providing these capabilities as a platform has two major advantages: first, the cost of development, deployment, and operations incurred by service developers is reduced as important capabilities can be outsourced to the platform; and, second, user experience is enhanced as users are presented with a cohesive environment in which "state" (e.g.., identities, groups, data) is shared between services.

The *Discovery Exchange* tackles the important challenge of sustainability. As described in §4.1.1, research software developers face the unenviable task of supporting their software in a short-term and novelty-oriented funding environment. New methods are required for supporting such

software, for example by distributing costs across those that benefit from usage. The Discovery Exchange will meet this need by providing the capabilities to support negotiation, management, and enforcement of subscriptions (with users, projects, or institutions) as well as the mechanisms for managing and collecting payments. Research services can in turn leverage these capabilities to augment existing funding streams with methods to recover development and operations costs from their users (or in many cases, the institutions to which these users belong).

6. The Discovery Cloud architecture

We next present our proposed Discovery Cloud architecture and explain how it must embrace, and as necessary extend, the modern, standard (or at least conventional) approaches being used to deliver services to other markets and to existing research cyberinfrastructure. We describe existing web and mobile standards that form the basis of the Discovery Cloud before categorizing and detailing the services that we believe are required to instantiate the Discovery Platform that will support the rapid development and delivery of powerful research services. We then discuss how research services can be delivered on this platform. Finally, we present an economic approach to sustainability using the Discovery Exchange.

6.1. Leveraging web and mobile standards

The small set of widely adopted standards and approaches that underpin the web will also be foundational to the Discovery Cloud. Services provided via the Discovery Cloud must be web-based services. Thus, the Discovery Cloud must deliver—and make it easy for others to deliver—web-based services for SMLs: services that can leverage, integrate with, and extend other web services both within and outside of our market.

As a general rule, we require a high degree of compatibility with web standards, so that Discovery Platform services integrate cleanly into the broader web. For example: HTTP, HTML, CSS, and JavaScript are foundational to anything web; TLS provides the standard security layer under HTTP (i.e., HTTPS), with X.509 server certificates providing for server authentication on those channels; OAuth2 [30] has emerged as the web standard for authorization of access to web resources; OpenID Connect [31], an extension to OAuth2, has emerged as the web standard for end-user identity verification; Most web services have adopted the HTTP-based, REST architectural style for interactions between clients and services; and JSON has emerged as the primary data-interchange format used by REST services on the web, due to its affinity with JavaScript.

Mobile devices have become important doorways to software and data systems, but currently see limited uptake for science applications. SaaS adoption also initially lagged in science but is now beginning to take off; we expect mobile applications to follow the same trajectory in SMLs.

We envision two general classes of mobile applications in research: 1) mobile apps that help researchers perform their day-to-day work (e.g., analyzing data, producing results and reports); 2) mobile apps that have specialized uses in applied research contexts (e.g., collecting data from experiments). In both cases, mobile support provides flexibility by not limiting the researcher to specific environments, and more importantly, it enables access to the rest of the larger infrastructure in settings where traditional platforms would not work, e.g. data capture in remote locations.

Leveraging existing web standards allows the Discovery Platform, and research services that are available via it, to use the myriad of networking technologies that have built up around these standards, such as load balancers, proxies, and content distribution networks. It also enables standard integration with existing services, platforms, and clients.

6.2. The Discovery Platform

PaaS makes it far easier and cheaper to create, maintain, and operate SaaS web, mobile, and desktop applications. In the commercial and consumer markets, virtually all new applications leverage commercial PaaS from AWS, Google, and Microsoft Azure. And because platform services use REST standards, they can be leveraged by other applications and services, regardless of the programming languages and frameworks used by those other applications and services.

However, commercial PaaS is not sufficient for the science market. Instead we believe that the key to unlocking an ecosystem of research services is to deliver the Discovery Platform. The platform builds upon and enhances commercial PaaS, by providing a set of platform services specifically attuned to the needs of science application developers. These services can then be used by scientific SaaS developers to improve the efficiency and ease of creating and delivering their own services to the research community. Important core platform services are those that are required by the vast majority of SaaS developers, for example authentication, identity, and group models, as well as data access and management capabilities. Figure 2 depicts the role of the Discovery Platform.

In Globus, we have developed and deployed services that are required by a large number of research users. Thus, these services can be used to satisfy the core capabilities of the Discovery Platform. However, because requirements are continually evolving, so too must the core services. We must therefore continually evaluate user requirements and and develop or adopt new services as required.

6.2.1. Core platform services. We observe in §4.1 that science applications are routinely built as silos, caused by each implementing core functionality (e.g., security) in different and incompatible ways. The unfortunate consequences of this siloed approach include increased costs to build, maintain, and operate each application, since there is often little reuse; decreased functionality, as most application developers do not have the resources or expertise to create rich core functionality; inconsistent user experience across applications; and an inability for applications to integrate with each other, due to incompatible core functionality.

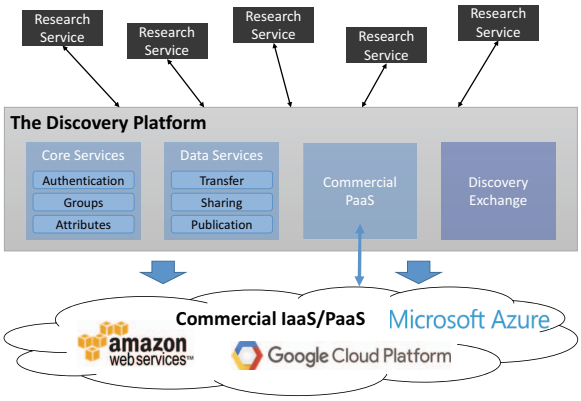

Figure 2. The Discovery Cloud. The Discovery Platform builds upon commercial IaaS and PaaS to provide a set of core services that can be leveraged and built upon by research services

In our work to date, we have identified the following core platform services as essential for a Discovery Platform that aims to make it easier and cheaper to create interoperable research services.

Globus Auth: This platform service provides authentication and authorization brokering between users, applications, services, and identity providers. The science ecosystem comprises thousands of independent identity providers and tens of thousands of independent resource/service providers. Researchers and their applications may want to use essentially any combination of these identity and resource/service providers. Globus Auth manages security interactions between all of these independent parties—not just user login to applications, but also application-to-service and service-to-service interactions. Independent software developers may use Globus Auth instead of developing their own authentication and authorization systems. Globus Auth integrates well with the national and international (federated) identity systems already in place in the scientific/research community, e.g., the InCommon federation [32], and with web standards such as OAuth2 and OpenID Connect.

Globus Groups: This platform service allows end users and application developers to define and manage groups that reflect the oftentimes complex and dynamic organizational structures of modern research collaborations. These groups can then be used by developers to control access to resources within their own systems.

Globus Attributes: This platform service allows end users and application developers to create, manage, and retrieve key/value attributes associated with identities, which can be used by developers to support sophisticated, attribute-based authorization features within their services.

A key benefit of these core services is that they are shared by all applications and services that use them. The same identities, groups, and attributes work across all applications and services that leverage the Core platform services, rather than each being its own silo, as is typically the case with science applications and services today. Users benefit by being able to move seamlessly through this ecosystem (e.g., they can define their project group once, and use it across the whole ecosystem of applications and services), and service developers benefit by being able to seamlessly integrate with other services.

6.2.2. Research data management platform services. Data is vital to both observational and simulation-based scientific methods. As computers, networks, and automated sensors have become more accessible to scientists, the amount of data generated by and used by scientific projects in many fields of study has exploded. Globus currently provides platform support for three data management challenges: transfer, sharing, and publication. Each data platform service is designed to be used by independent scientific software developers in their own applications, reducing the overall cost of creating and maintaining applications to satisfy specialized scientific needs.

Globus transfer: The need to move data from one location to another is common to many scientific enterprises, particularly when inter-institutional collaboration is required or when regional, national, or international resources (HPC, large-scale instruments, cloud) are being used. Globus Transfer provides a platform that makes it easy for developers to include high-speed, secure, and highly reliable data movement features in their applications.

Globus sharing: Scientific collaboration typically requires shared access to data, but, as in all human enterprises, there is usually a twist or two in how that sharing is conducted. Access control can be especially important in scientific collaboration, so Globus Sharing provides a platform that enables scientific software developers to add easy-to-use data sharing features and these all-important sharing controls to their applications.

Globus publication: Publishing results is a vital phase of science. Increasingly, scientific results include data and other digital artifacts. Globus Publication provides a platform that enables scientific software developers to add data publication features to their applications.

6.2.3. Commercial PaaS gateways. The Discovery Platform aims to extend and adapt commercial PaaS (e.g., Amazon Web Services, Google Cloud Platform, Microsoft Azure) for science, not replace it. However, in order to make commercial PaaS more accessible to science applications, the Discovery Platform must include various gateway services to commercial PaaS.

Security gateways: Each commercial platform has its own PaaS security approaches. Discovery Platform core services (e.g., Globus Auth, Globus Groups) already allow for simplified integration with commercial PaaS security.

Data gateways: Commercial data analytics platform services can be of use to the science community: for example, AWS Redshift, AWS Elastic MapReduce, and AWS Machine Learning. Discovery Platform research data services must integrate with such commercial platform services, so that they can be used seamlessly as part of the platform.

6.3. Software as a Service for research

A multitude of research services can be developed and deployed on top of the Discovery Platform. In fact, most research software that exists today could be migrated to such a model. Research services could support any phase of the research lifecycle from data acquisition through to publication; may be aimed at any group of users (e.g., the entire research community, specific institutions, research projects, researchers, postdocs, students, etc.); and be built upon any combination of platform services and other research services. We briefly describe potential services that relate to data, analysis, and collaborative activities.

Data: Many different data repositories serve various segments of the research community (e.g., genetic data, medical images, climate data, policy models, materials properties, etc.). These repositories are currently offered in various formats and without any form of integration or interoperation between repositories. If developed on top of the Discovery Platform these services would benefit from the use of common user identities and attributes (e.g., researcher's institution, project, PI, etc.), high performance data access methods for downloading or depositing potentially large datasets, and the ability to more easily integrate related repositories. Similarly, methods for storing and managing structured data (e.g., SQLShare, ERMrest [33]) could benefit from platform methods for referencing remotely accessible data and enforcing authorizations derived from service-independent group memberships.

Analysis: The most valuable output of research is often new models and analysis methods that can be applied or built upon by other researchers. Researchers repeatedly invent new ways to compare genomes, analyze the structure of materials, and model the effect of greenhouse gases. But before other researchers can run these new programs, either the developer or user must gain familiarity with and instantiate a whole machinery for providing access to models, supplying input data, scheduling job submission and execution requests, and allocating the cyberinfrastructure—all capabilities that can be outsourced to the Discovery Platform. Examples of services that perform such tasks include Galaxy and NanoHUB for executing arbitrary workflows and applications on behalf of their users.

Collaboration: A third class of service to be included in the Discovery Cloud are those designed to support the collaborative research process: for example, collaborative document editing, wikis for recording and discussing ideas, software versioning repositories, data sharing software, and electronic lab notebooks. While many commercial entities offer services in this space, those services can still benefit from Discovery Platform capabilities. For example, they can be extended to allow researchers to use a single common identity that can be linked across services, to leverage group-based authorization for managing access to a group's resources, and to provide common methods for accessing and sharing data associated with multiple services.

7. Applying the Discovery Platform

We use two examples from our Globus work to elucidate how the Discovery Platform can support developers of Discovery Cloud services.

The US Department of Energy's Systems Biology Knowledge Base (KBase: http://kbase.us), is a web application that provides "an open platform for comparative functional genomics and systems biology for microbes, plants and their communities, and for sharing results and methods with other scientists." Recognizing that they needed sophisticated federated identity and group management capabilities similar to those provided in the Globus web application, the kBase group adopted the Globus Auth platform services, allowing them to focus their scarce resources on the scientific value-add of their applications rather than on plumbing.

The CISL Research Data Archive (RDA: http://rda.ncar.edu), "contains a large and diverse collection of meteorological and oceanographic observations, operational and reanalysis model outputs, and remote sensing datasets to support atmospheric and geosciences research, along with ancillary datasets, such as topography/bathymetry, vegetation, and land use." Tens of thousands of users login to the RDA web application, search for data within collections, and download that data for use in their own research. Historically, CISL would extract the requested data for the user into a tar ball, and make it available for download via HTTP from the RDA web site. That approach became increasingly problematic as dataset sizes grew to terabytes scales. To address this challenge, the RDA team integrated with Globus Auth and Globus Groups platform services so that RDA users can use Globus file transfer capabilities to more easily and quickly download data sets.

In both cases, Discovery Platform services allowed research application developers to deliver sophisticated functionality easily, cheaply, and reliably. These services will be complemented by services, such as the Discovery Exchange, that will allow research application developers to provide services to users via freemium models.

8. The Discovery Exchange

The development and support of any software requires funding, whether in the form of volunteer effort, grants from agencies or foundations, or directly by users. SaaS approaches reduce but do not eliminate the need for funding. Thus, a major obstacle to sustainable research services is inevitably the mechanics of collecting subscription revenue, negotiating contracts, accepting payments, marketing new capabilities, etc. To address this need, we propose the *Discovery Exchange* as a single point of contact for institutions, projects, and individuals wanting to subscribe to superior research services. It will include a services registry, but the important and hard part will be the business side: marketing, market research, legal, financial, etc. It will also provide market research services to potential providers, e.g., by polling potential customers.

Core Discovery Exchange capabilities will include directory/information services to keep track of available offerings; quality control services to ensure that the system is functioning and available; payment models that service providers can leverage; and a support system for providing cohesive customer support across services.

The Discovery Exchange provides an opportunity to explore alternative payment models, usage models, and currencies for the research community. Commercial service providers have adopted a range of freemium, subscription, and usage-based models. In Globus we have also explored these models with varying degrees of success, as discussed in §9. The Discovery Exchange must have the flexibility to enable the use of both existing models as well as new models by the Discovery Platform and also by participating research services. Traditional money is, of course, a useful currency, but we know that the source is often indirect: the host institution rather than the researcher (or the researcher's assistants) working in their SML. There may be other useful currencies, such as paying service providers in free cloud cycles or data providers in free storage. Citations could be of value for some data and service providers.

9. Addressing the sustainability challenge

Sustainability is a challenge for all research service (and application) developers. Pricing and available payment mechanisms do not match up with small-scale research realities; services are designed with different styles of research in mind (e.g., large-scale collaborations with plenty of available human capital); services are designed, deployed, and operated for limited-time grand challenge activities and are intended to be decommissioned once the initial purpose is satisfied; and specialized research services are not attractive to business, which seeks quick payoffs and easy, low-cost sales opportunities. We introduce the Discovery Exchange to provide sustainability mechanisms for the Discovery Platform and for the Exchange itself.

In our work with Globus, we have explored various freemium approaches to sustainability. We first focused on converting individual researchers from free users to paying subscribers, offering the Globus file transfer service free of charge and requiring a subscription to enable file sharing and other advanced features. We encountered a number of challenges under this approach, primarily the disproportionately high effort required to manage individual subscription payments and the reluctance of individual researchers to subscribe to services. We observed that researchers prefer to use capabilities offered by their campus computing department rather than to pay external providers for such capabilities. Thus, we shifted our focus to enlisting institutional subscribers. We are now starting to see confirmation that the SaaS approach is viable for delivering sustainable research cyberinfrastructure, as supported by the willingness of more than 30 institutions to subscribe to premium services within the first 18 months of launch.

While encouraged by the initial success of this approach, we encountered challenges that must be overcome for the Discovery Cloud to be sustainable over the long term. For example, many institutions have (widely varying) contractual requirements that constrain their ability to deploy new (or alternative) SaaS solutions: requirements often rooted in more traditional purchases (i.e., licensed software) from larger vendors. Smaller vendors (typical of many SaaS providers) are unable to accept the burden of some terms, and the time-consuming process of negotiating with each institution individually adds substantially to the cost of operating a SaaS enterprise. To mitigate such concerns, the Discovery Cloud will thus promote a standardized set of terms, based upon our initial experiences with early Globus subscribers, that would apply to a broad variety of services. Thus, services using the Discovery Exchange would be able to implement advanced subscription models without needing to establish complex agreements with individual subscribers.

10. Summary

We are convinced that the *Discovery Cloud* represents the future of scientific computing. Once realized, it will allow any researcher, in any laboratory, to access, via intuitive interfaces, a rich set of services that collectively automate and accelerate common research activities. Researchers working within SMLs will be able to discover any computational, software, or data resource relevant to their research; track and organize data consumed and produced by their research; access and run powerful modeling and simulation software; and collaborate with colleagues regardless of location—all without installing software, acquiring storage systems or computational infrastructure, or employing IT staff to operate and maintain hardware and software. The Discovery Cloud will thus transform research practice across SMLs worldwide, in much the same way that the cloud has transformed business practice.

Key to the power of the Discovery Cloud is its use of cloud-hosted services to reduce cost and complexity. Industry experience shows that SaaS can reduce both marginal cost of delivery and cost of ownership to customers. In industry, providers leverage these benefits to maximize revenue, margins, and/or profits. The Discovery Cloud can instead focus on maximizing value to the entire science community while achieving sustainability. Integrating costs incurred across Discovery Cloud providers and consumers, the result can be a dramatic reduction in total IT costs and/or an equally dramatic increase in delivered capability.

While the vision of the Discovery Cloud may appear unrealistic, we have presented a body of evidence that suggests scientists are in need of such capabilities and are willing to use them if offered access. A key to success, we argue, is the *Discovery Platform*, which leverages core services (authentication, groups, data management), many of which we have already developed, deployed, and operated for several years, and many of which are now relied on by thousands of researchers. Over time, these core services will be extended and adapted to meet the needs of a broader set of researchers and research services, with the crucial

Discovery Exchange supporting the broad development and operation of sustainable scientific services.

We are hard at work on the realization of this model. We are developing, enhancing, and deploying core platform services that can be leveraged by other developers. Simultaneously, we are working with external groups to apply these platform services in production settings. We will next develop new core services and adapt and integrate other capabilities, including those originally developed for enterprise and big science purposes.

Acknowledgments

We thank the Globus team for implementing and operating Globus services, and the users of those services for their continued support. This research was supported in part by DOE contract DE-AC02-06CH11357; NIH contract 1U54EB020406-01, Big Data for Discovery Science Center; and NIST contract 60NANB15D077.

References

[1] F. Chong and G. Carraro, "Architecture strategies for catching the long tail," *MSDN Library, Microsoft Corporation*, pp. 9–10, 2006.

[2] A. Dubey and D. Wagle, "Delivering software as a service," *The McKinsey Quarterly*, vol. May, 2007.

[3] I. Foster, V. Vasiliadis, and S. Tuecke, "Software as a service as a path to software sustainability," tech. rep., 2013. http://dx.doi.org/10.6084/m9.figshare.791604.

[4] G. Lawton, "Developing software online with platform-as-a-service technology," *Computer*, vol. 41, no. 6, pp. 13–15, 2008.

[5] I. Foster, "Globus Online: Accelerating and democratizing science through cloud-based services," *IEEE Internet Computing*, vol. 15, no. 3, pp. 70–73, 2011.

[6] P. B. Heidorn, "Shedding light on the dark data in the long tail of science," *Library Trends*, vol. 57, no. 2, pp. 280–299, 2008.

[7] N. Weber, "IDCC 2013 Poster: The tail and the telling - Distributions of NSF funding and long-tail data." http://dx.doi.org/10.6084/m9.figshare.106458, 01 2013.

[8] "NIH Data Book." https://report.nih.gov/NIHDatabook/. Web site. Accessed: Jan 21, 2016.

[9] N. S. Board, "The Science and Engineering Workforce: Realizing America's Potential," Tech. Rep. NSF0369, National Science Foundation, Arlington, VA, 2003.

[10] J.-M. Fortin and D. J. Currie, "Big science vs. little science: How scientific impact scales with funding," *PLoS ONE*, vol. 8, pp. e65263+, June 2013.

[11] T. Hey, S. Tansley, and K. Tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.

[12] Data2Insight, "University of Washington data science environment: Software tools, environments, and support final evaluation findings," tech. rep., University of Washington, 2014.

[13] T. Dasu and T. Johnson, *Exploratory data mining and data cleaning*, vol. 479. John Wiley & Sons, 2003.

[14] S. Lohr, "For big-data scientists, 'janitor work' is key hurdle to insights," *New York Times*, August 17 2014.

[15] S. A. Goff, M. Vaughn, S. McKay, E. Lyons, A. E. Stapleton, D. Gessler, N. Matasci, L. Wang, M. Hanlon, A. Lenards, *et al.*, "The iPlant collaborative: Cyberinfrastructure for plant biology," *Frontiers in Plant Science*, vol. 2, 2011.

[16] G. Klimeck, M. McLennan, S. P. Brophy, G. B. Adams III, and M. S. Lundstrom, "nanohub.org: Advancing education and research in nanotechnology," *Computing in Science & Engineering*, vol. 10, no. 5, pp. 17–23, 2008.

[17] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, and N. Wilkins-Diehr, "XSEDE: Accelerating scientific discovery," *Computing in Science and Engineering*, vol. 16, no. 5, pp. 62–74, 2014.

[18] R. Pordes, D. Petravick, B. Kramer, D. Olson, M. Livny, A. Roy, P. Avery, K. Blackburn, T. Wenaus, F. Würthwein, *et al.*, "The Open Science Grid," *Journal of Physics: Conference Series*, vol. 78, no. 1, p. 012057, 2007.

[19] W. Michener, D. Vieglais, T. Vision, J. Kunze, P. Cruse, and G. Janée, "DataONE: Data observation network for earth-preserving data and enabling innovation in the biological and environmental sciences," *D-Lib Magazine*, vol. 17, p. 3, 2011.

[20] H. Karasti, K. S. Baker, and E. Halkola, "Enriching the notion of data curation in e-science: Data managing and information infrastructuring in the long term ecological research (LTER) network," *Computer Supported Cooperative Work (CSCW)*, vol. 15, no. 4, pp. 321–358, 2006.

[21] D. Lifka, I. Foster, S. Mehringer, M. Parashar, P. Redfern, C. Stewart, and S. Tuecke, "XSEDE cloud survey report," 2013. http://hdl.handle.net/2142/45766.

[22] J. Goecks, A. Nekrutenko, J. Taylor, and T. G. Team, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences," *Genome Biol*, vol. 11, no. 8, p. R86, 2010.

[23] B. Howe, G. Cole, E. Souroush, P. Koutris, A. Key, N. Khoussainova, and L. Battle, "Database-as-a-service for long-tail science," in *Scientific and Statistical Database Management*, pp. 480–489, Springer, 2011.

[24] S. Wuchty, B. Jones, and B. Uzzi, "The increasing dominance of teams in production of knowledge," *Science*, vol. 316, pp. 1036–1039, 18. Mai 2007 2007.

[25] E. Yakel, I. M. Faniel, A. Kriesberg, and A. Yoon, "Trust in digital repositories," *IJDC*, vol. 8, no. 1, pp. 143–156, 2013.

[26] K. Chard, M. Lidman, B. McCollam, J. Bryan, R. Ananthakrishnan, S. Tuecke, and I. Foster, "Globus Nexus: A platform-as-a-service provider of research identity, profile, and group management," *Future Generation Computer Systems*, vol. 56, pp. 571–583, 2016.

[27] B. Allen, J. Bresnahan, L. Childers, I. Foster, G. Kandaswamy, R. Kettimuthu, J. Kordas, M. Link, S. Martin, K. Pickett, and S. Tuecke, "Software as a service for data scientists," *Communications of the ACM*, vol. 55, no. 2, pp. 81–88, 2012.

[28] K. Chard, J. Pruyne, B. Blaiszik, R. Ananthakrishnan, S. Tuecke, and I. Foster, "Globus data publication as a service: Lowering barriers to reproducible science," in *11th IEEE International Conference on eScience*, 2015.

[29] R. K. Madduri, D. Sulakhe, L. Lacinski, B. Liu, A. Rodriguez, K. Chard, U. J. Dave, and I. T. Foster, "Experiences building Globus Genomics: A next-generation sequencing analysis service using Galaxy, Globus, and Amazon Web Services," *Concurrency and Computation: Practice and Experience*, vol. 26, no. 13, pp. 2266–2279, 2014.

[30] D. Hardt, "OAuth 2.0 authorization framework specification," 2012. http://tools.ietf.org/html/rfc6749 [accessed May 1, 2014].

[31] N. Sakimura, J. Bradley, M. Jones, B. d. Medeiros, and C. Mortimore, "OpenID Connect Core 1.0 incorporating errata set 1," 2014. http://openid.net/specs/openid-connect-core-1_0.html.

[32] W. Barnett, V. Welch, A. Walsh, and C. A. Stewart, "A roadmap for using NSF cyberinfrastructure with InCommon," 2011.

[33] R. Schuler, C. Kesselman, and K. Czajkowski, "Data centric discovery with a data-oriented architecture," in *Proceedings of the 1st Workshop on The Science of Cyberinfrastructure: Research, Experience, Applications and Models (SCREAM)*, pp. 37–44, 2015.