HPC4Health

The HPC4Health Network

Building an Ontario-wide platform for human health data

HPC4Health: High-Performance Data and Computing for Health Care

PUBLISHED BY HPC4HEALTH

WWW.HPC4HEALTH.CA

Executive Summary

The explosion of human health data can mean new advances in human health care — but only if the infrastructure, architecture, policies, and expertise are brought together in innovative ways to make effective, safe research use of the data. We have such an example of a collaborative center in Ontario, in HPC4Health.

An innovative and successful shared platform for health genomics data and analysis, the original incarnation of HPC4Health in 2014 was a small pilot project between SickKids and UHN; today, the collaboration counts amongst its employees six scientific and eight technical staff, works with four partner institutions, and provides 7,000 compute cores and over 2 petabytes of secure data storage.

With continued success both technical and scientific HPC4Health faces a question — how best to expand to meet growing demand in institutions across Ontario. Should the organization meet growing by increasing the capacity of its current operations, or by developing more such operations across the province, and connecting them all into a coherent whole?

While current facilities have room to grow, there are several factors that will limit scaling HPC4Health as a centralized resource over time. As with any data-intensive science, the need to keep compute near the modern high-throughput data generation facilties, and the desire to make use of local clusters of expertise and resources, both suggest against single, centralized facilities. And in the particular case of human health, data governance policies may require a more decentralized approach.

The H4H Network, with pilot projects already in place in Montréal and Hamilton, will build a coherent, shared network of resources and expertise available to health researchers across Ontario. With each site modelled after the current HPC4Health site at SickKids, the network will use: a single, coherent security policy; a common software stack; and a federated network of genomic experts, all presenting a common set of services to researchers while providing international best-practices in security, privacy, and governance. The H4H Network will unlock genomic data and local expertise for benefit of all Ontarians, while maintaining the security and privacy of all patient data.

Human Health Research in the Era of Genomics

Next-generation genomics and electronic medical records have become ubiquitous almost simultaneously, opening completely new windows onto the field of human health research.

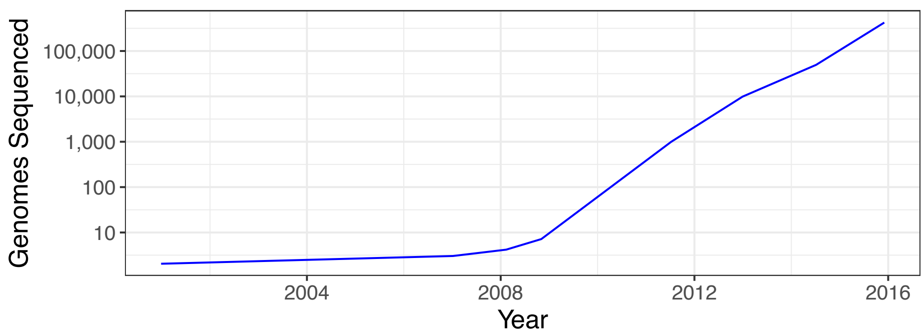

Figure 1: The cumulative number of human genomes sequenced over the past 15 years, data from [Stephens et al., 2015]. New technologies and data types have caused a inflection in the exponential rate of data growth which continues today.

In genomics, an exponentially-growing number of human genomes are being sequenced, making enormous amounts of evidence on the heredity of human disease states at previously inaccessible levels of detail (Fig 1). Data rates are only increasing as new devices become available, and new types of data are being generated; for instance, RNA sequencing allows us to go beyond simply sequencing "the" genome of an individual and instead measure the gene products being expressed in particular cells at particular times, giving us insight into not just predispositions but the precise disease state of cells over time.

Until now, the bulk of this genomic data has come from research projects — focused, generally short-term efforts to answer specific questions using genomic information. However, two changes in the practice of medicine are poised to radically increase the rate of genomic and human health data creation in Canada.

First, genomic medicine[1] is beginning to enter the standard of care, initially in the cases of hard-to-diagnose rare diseases or recurrent cancers, and starting to displace smaller and more limited genetic tests in other areas. The sheer scale of clinical medicine — in Canada, hospital spending alone is

[1] Genomic Medicine, as defined by the NIH's National Human Genome Research Institute: "An emerging medical discipline that involves using genomic information about an individual as part of their clinical care (*e.g.*, for diagnostic or therapeutic decision-making)"

nearly seventy times the research funding budget of the CIHR — means that as adoption increases, clinical genomic data creation will rapidly outpace that of research genomic data.

Second, information technology advances elsewhere in healthcare has led to the rapid adoption of Electronic Medical Record systems[2], describing a patient's condition, tests, and treatment in detail and at least partially in machine-readable form.

[2] Electronic Medical Records, or EMRs, are digital version of paper chapters in hospitals or doctors offices; Electronic Health Records, or EHRs, are more integrated systems combining information across practices and institutions. EMRs are mature and growing in adoption rapidly, while EHR systems are still some time from being common.

The rapid growth of genomic data volumes and the increasing depth and detail of clinical data present in EMRs offers enormous promise for human health research, with insights into both basic biology and to future treatments. The joint analyses of clinical and phenotypic health record data along with genomic information about the patient and their disease offers unprecented opportunity for researchers to connect genetic predisposition, treatments, and outcomes, allowing the development of national, truly precision, medicine practices.

But making use of this data in an era of rapid growth raises multiple challenges. On the physical infrastructure side, simply making available the storage resources to capture and archive the onrush of data is a daunting effort, along with providing the computational power to perform increasingly sophisticated analyses. Architecting, building, and maintaining these systems, particularly tuned to the needs of health research, requires a specialized approach.

Human infrastructure is also required. Making productive use of the data means ensuring that the expertise exists and is available to for interpretation, that those experts are continuously kept up to date on the new types of data and new techniques for analysis, and have time to develop novel methods to address the questions they are tasked with answering. This requires funding, ongoing training, and opportunities for professional recognition and growth of the experts, whether they be bioinformaticians, computational biologists, or systems administrators.

Finally, the unique challenges of dealing with health data means that the duty of care to patients to zealously protect the security and their privacy is paramount; sophisticated and enforcable policies around data governance and consent are required, along with international best practices around security and monitoring.

The explosion of human health data can mean new advances in human health care — but only if the infrastructure, architecture, policies, and expertise are brought together in innovative ways to make effective, safe research use of the data. We have such an example of a collaborative center in Ontario, in HPC4Health.

HPC4Health — a Made-in-Ontario Approach

As early and major players in genomic research and medicine, in 2014 SickKids and the University Health Network (UHN) faced a problem — how to manage and make use of the influx of genomic and other health data they were already charged with storing and analyzing.

With a common challenge, and building on existing partnerships, the institutions formed a first-of-its-kind collaboration and in Ontario, combining forces and sharing resources to build HPC4Health[3]: a shared-services approach, building a cross-institution center of infrastructure and expertise for the analysis of human health and genomic data. This innovative partnership gathered much attention in the hospital, genomics, and research computing communities [Hospital News, 2014, Genome Web, 2014, Inside HPC, 2015, Telfer, 2016].

[3] http://www.hpc4health.ca

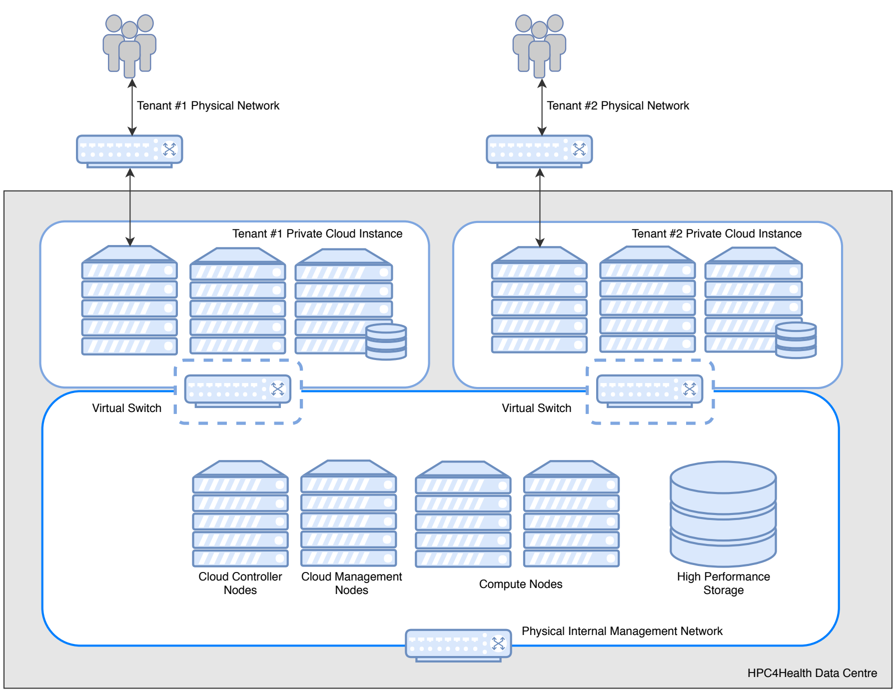

Figure 2: The HPC4Health Architecture. An on-premesis (at SickKids) private cloud is partitioned into tenant-controlled secure data environments for the institutions, with shared administrative resources managed by the HPC4Health administrators. Tenants can get technical support from the core technical staff, or scientific support from the growing team of genomics, machine learning, and health record experts.

HPC4Health's compute and storage infrastructure (Fig 2) consists of an elastic secure cloud — an arrangement that functions like an office building rented to multiple tentants. The "superintendant" maintains core shared facilities, but the "tenants" have complete control of their own office space. This is implemented through in an on-premises private OpenStack[4] cloud, partitioned securely into an administrative core and "tenant" environments.

[4] https://www.openstack.org

Crucially, HPC4Health is more than computational infrastructure; it is also experience and expertise. HPC4Health is governed by a sophisticated policy and data governance framework; it is governed by an executive committee and a science advisory panel; and is run by teams of technical and scientific personnel that every day make sure that health insight is successfully distilled from health data.

The original incarnation of HPC4Health in 2014 was a small pilot project between SickKids and UHN; today, the collaboration counts amongst its employees six scientific and eight technical staff, works with four partner institutions, and provides 7,000 compute cores and over 2 petabytes[5] of secure data storage from its primary data centre in SickKids research institute building.

[5] A petabyte is over a million gigabytes, and is enough to store over three years worth of 24/7 high-definition video, or raw data from nearly 7,500 whole genome sequencing experiments.

With continued success both technical and scientific (with over XXX **TODO** publications), HPC4Health faces a question — how best to expand to meet growing demand in institutions across Ontario. Broadly, should HPC4Health meet demand by scaling vertically or horizontally[6]; should HPC4Health serve more researchers by increasing the capacity of its current operations, or by developing more such operations across the province, and connecting them all into a coherent whole?

[6] Scaling a system *vertically*, or "scaling up", means making the components of a system more powerful — bigger, faster, greater capacity — while scaling *horizontally*, or "scaling out", means increasing the number of components. The benefits of horizontal scaling when appliciable are generally flexibility, cost-effectiveness, and fault-tolerance, while that of vertical scaling are typcially simplicity.

While HPC4Health operations has room to grow, there are several factors that will limit the feasibility of vertical scaling over time; many of these are common to data-intensive science generally:

Keeping compute near the data generation: Data generation rates of modern devices are growing much more quickly than network speeds. Individual modern sequencing devices can output data in excess of 1,000 gigabytes per hour, straining or saturating wide-area network links.

Making use of local resources and expertise: Significant investments have been made across Ontario in training of both highly qualified personnel, often in clusters with significant expertise in particular problems or analyses types, and in equipment. Making effective use of those resources is crucial to cost-effective expansion.

and one particular to human health data:

Data governance policies: Many stewards of human health data have policies in place about keeping data on-premesis or limiting where the data can go (such as going to another jurisdiction) or who can see it. In this case, building a local operation may be required.

The H4H Network — Building on Strengths

To build on HPC4Health successes and strengths and investments elsewhere, we take a system-wide approach with the H4H Network — allowing the benefits of coordation such as economies of scale from consistent design, shared procurement, unified management; consistent services for researchers; and best-practice security adminstration — while also taking advantage of local clusters of expertise and resources, and placing facilites near data generation.

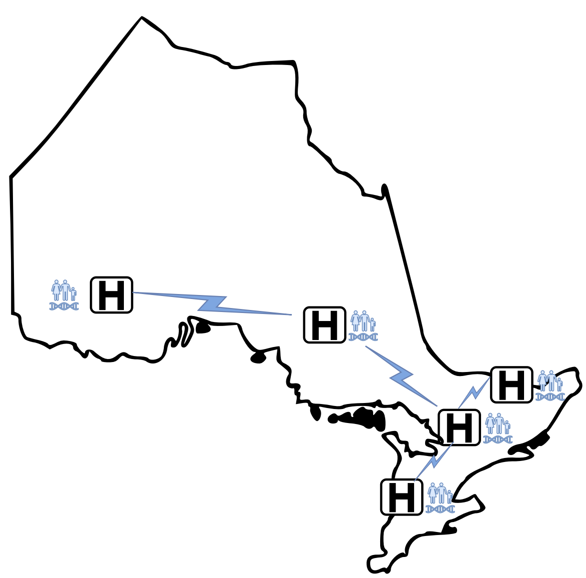

Figure 3: The H4H Network, with pilot projects already in place in Montréal and Hamilton, will build a coherent, shared network of resources and expertise available to health researchers across Ontario. With each site modelled after the current HPC4Health site at SickKids, the network will use: a single, coherent security policy; a common software stack; and a federated network of genomic experts, all presenting a consistent services to researchers while providing international best-practices in security, privacy, and governance.

Pilot projects implementing systems modelled after HPC4Health — within Ontario, in Hamilton, as well as out of province with the McGill secure health cloud — show the feasibility exporting the architecture designed for the original HPC4Health. But for researchers to make the best possible use of the health data entrusted to them by patients, and to take advantage of scale in purchasing, administration, and security, it is not enough to have Ontario's genomic research data isolated in silos. These sites must be part of a connected network, with researchers able to access the combined expertise — and, when appropriately authorized, particticular data sets — from across the network.

The next step for the H4H Network is, starting with existing pilots, to build that secure network of human health data sites, with coherent policies, security domain and monitoring, and with data available for analysis by researchers using existing projects and services such as GenAP[7] and CanDIG[8].

We propose to begin by expanding the Hamilton site and unifying it with common security domain, operational procedures, and data governance policies; and once those policies have been successfully developed to encompass multiple sites, expanding to a third Ontario site at one of several possible institutes.

[7] genap.ca

[8] distributedgenomics.ca

Bibliography

Genome Web. Sickkids, UHN pilot new cloud infrastructure for healthcare. *Genome Web*, Nov 2014. URL https://www.genomeweb.com/informatics/sickkids-uhn-pilot-new-cloud-infrastructure-healthcare.

Hospital News. Sickkids-UHN partnership brings cloud technology to health care. *Hospital News*, Nov 2014. URL http://hospitalnews.com/sickkids-uhn-partnership-brings-cloud-technology-health-care/.

Inside HPC. Hpc4health selects mellanox infiniband for cancer and genomics research. *Inside HPC*, Sept 2015. URL https://insidehpc.com/2015/09/hpc4health-selects-mellanox-infiniband-for-cancer-and-genomics-research/.

Zachary D Stephens, Skylar Y Lee, Faraz Faghri, Roy H Campbell, Chengxiang Zhai, Miles J Efron, Ravishankar Iyer, Michael C Schatz, Saurabh Sinha, and Gene E Robinson. Big data: astronomical or genomical? *PLoS biology*, 13(7):e1002195, 2015. URL https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002195.

Stig Telfer. The crossroads of cloud and HPC: Openstack for scientific research. Technical report, StackHPC Ltd, October 2016. URL https://www.openstack.org/assets/science/OpenStack-CloudandHPC6x9Booklet-v4-online.pdf.