Optional!

If you choose to work on exercises, or leave, you will not feel uninformed for tomorrow's work.

We avoid platform-specific information as much as possible, but some of you do want to hear about an actual enterprise class environment, and real applications. We briefly do that here.

John Urbanic, Parallel Computing Scientist
The Shift to Big Data

**New Emphases**

- **Social networks and the Internet**
- **Video**
- **Collections**
- **Legacy documents**
- **Environmental sensors**: Water temperature profiles from tagged hooded seals

**Pan-STARRS telescope**
http://pan-starrs.ifa.hawaii.edu/public/

**Genome sequencers** (Wikipedia Commons)

**NOAA climate modeling**
http://www.oml.gov/info/omlreview/v42_3_09/article02.shtml

**Horniman museum**: http://www.horniman.ac.uk/get_involved/blog/bioblitz-insects-reviewed

**Library of Congress stacks**
https://www.flickr.com/photos/danlem2001/6922113091/

**Wikipedia Commons**
Challenges and Software are Co-Evolving

- Structured Data
- Statistics
- Optimization (numerical)
- Calculations on Data
- Scientific Visualization
- Unstructured Data
- Machine Learning
- Optimization (decision-making)
- Natural Language Processing
- Video
- Sound
- Information Visualization
- Graph Analytics
- Image Analysis
Motivating Use Cases

Data-intensive applications & workflows
Gateways – the power of HPC without the programming
Shared data collections & analyses: cross-domain analytics
Deep learning
Graph analytics, machine learning, genome sequence assembly, and other large-memory applications
Scaling beyond the laptop
Scaling research to teams and collaborations
In-memory databases
Optimization & parameter sweeps
Distributed & service-oriented architectures
Data assimilation from large instruments & Internet
Leveraging an extensive software collection

Research areas that haven’t used HPC
Nontraditional HPC approaches to fields such as the physical sciences
Coupling applications in novel ways
Leveraging large memory and high bandwidth
Gateways and Tools for Building Them

Gateways provide easy-to-use access to Bridges’ HPC and data resources, allowing users to launch jobs, orchestrate complex workflows, and manage data from their browsers.

- **Extensive leveraging of databases and polystore systems**
- **Great attention to HCI is needed to get these right**

Interactive pipeline creation in GenePattern (Broad Institute)

Col*Fusion portal for the systematic accumulation, integration, and utilization of historical data, from http://colfusion.exp.sis.pitt.edu/colfusion/

Download sites for MEGA-6 (Molecular Evolutionary Genetic Analysis), from www.megasoftware.net
Virtualization and Containers

• Virtual Machines (VMs) enable flexibility, security, customization, reproducibility, ease of use, and interoperability with other services.

• User demand is for custom database and web server installations to develop data-intensive, distributed applications and containers for custom software stacks and portability.

• Bridges leverages OpenStack to provision resources, between interactive, batch, Hadoop, and VM uses.
High-Productivity Programming

Supporting languages that communities already use is vital for them to apply HPC to their research questions.
Bridges’ large memory is great for Spark!

*Bridges* enables workflows that integrate Spark/Hadoop, HPC, and/or shared-memory components.
Deep Learning Frameworks on *Bridges*

- Caffe
- PyTorch
- TensorFlow
- Theano
- K
- NVIDIA DIGITS
Purpose-built Intel® Omni-Path
Architecture topology for data-intensive HPC

Bridges Virtual Tour:
https://www.psc.edu/bvt
Example: Causal Discovery Portal
Center for Causal Discovery, an NIH Big Data to Knowledge Center of Excellence

Browser-based UI
- Prepare and upload data
- Run causal discovery algorithms
- Visualize results

Web node
- Execute causal discovery algorithms
  - Apache Tomcat Messaging

Database node
- Authentication
- Data
- Provenance
- MySQL
- Other DBs

Pylon filesystem
- TCGA fMRI

Analytics:
- FGS and other algorithms, building on TETRAD

LSM Node (3TB)
- ESM Node (12TB)
- Memory-resident datasets
Looking towards tomorrow: GPU Nodes

*Bridges’* Phase 1 and Phase 2 nodes accelerate both deep learning and simulation codes:

**Phase 1:** 16 nodes, each with:
- **2 x NVIDIA Tesla K80 GPUs (32 total)**
- 2 x Intel Xeon E5-2695 v3 (14c, 2.3/3.3 GHz)
- 128GB DDR4-2133 RAM

**Phase 2:** +32 nodes, each with:
- **2 x NVIDIA Tesla P100 GPUs (64 total)**
- 2 x Intel Xeon E5-2683 v4 (16c, 2.1/3.0 GHz)
- 128GB DDR4-2400 RAM

**Kepler architecture**
- 2496 CUDA cores (128/SM)
- 7.08B transistors on 561mm$^2$ die (28nm)
- 2x24 GB GDDR5; 2x240.6 GB/s
- 562 MHz base – 876 MHz boost
- 2.91 Tf/s (64b), 8.73 Tf/s (32b)

**Pascal architecture**
- 3584 CUDA cores (64/SM)
- 15.3B transistors on 610mm$^2$ die (16nm)
- 16GB CoWoS® HBM2 at 720 GB/s w/ ECC
- 1126 MHz base – 1303 MHz boost
- 4.7 Tf/s (64b), 9.3 Tf/s (32b), 18.7 Tf/s (16b)
- Page migration engine improves unified memory
- 64 P100 GPUs → 600 Tf/s (32b)
<table>
<thead>
<tr>
<th>Type</th>
<th>RAM</th>
<th>#</th>
<th>CPU / GPU / SSD</th>
<th>Server</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESM</td>
<td>12 TB</td>
<td>2</td>
<td>16 × Intel Xeon E7-8880 v3 (18c, 2.3/3.1 GHz, 45MB LLC)</td>
<td>HPE Integrity Superdome X</td>
</tr>
<tr>
<td></td>
<td>12 TB</td>
<td>2</td>
<td>16 × Intel Xeon E7-8880 v4 (22c, 2.2/3.3 GHz, 55MB LLC)</td>
<td></td>
</tr>
<tr>
<td>LSM</td>
<td>3 TB</td>
<td>8</td>
<td>4 × Intel Xeon E7-8860 v3 (16c, 2.2/3.2 GHz, 40 MB LLC)</td>
<td>HPE ProLiant DL580</td>
</tr>
<tr>
<td></td>
<td>3 TB</td>
<td>34</td>
<td>4 × Intel Xeon E7-8870 v4 (20c, 2.1/3.0 GHz, 50 MB LLC)</td>
<td></td>
</tr>
<tr>
<td>RSM</td>
<td>128 GB</td>
<td>752</td>
<td>2 × Intel Xeon E5-2695 v3 (14c, 2.3/3.3 GHz, 35MB LLC)</td>
<td>HPE Apollo 2000</td>
</tr>
<tr>
<td>RSM-GPU</td>
<td>128 GB</td>
<td>16</td>
<td>2 × Intel Xeon E5-2695 v3 + 2 × NVIDIA Tesla K80</td>
<td></td>
</tr>
<tr>
<td></td>
<td>128 GB</td>
<td>32</td>
<td>2 × Intel Xeon E5-2683 v4 (16c, 2.1/3.0 GHz, 40MB LLC) + 2 × NVIDIA Tesla P100</td>
<td></td>
</tr>
<tr>
<td>GPU-AI16</td>
<td>1.5 TB</td>
<td>1</td>
<td>16 × NVIDIA V100 32GB SXM2 + 2 × Intel Xeon Platinum 8168 + 8 × 3.84 TB NVMe SSDs</td>
<td>NVIDIA DGX-2 delivered by HPE</td>
</tr>
<tr>
<td>GPU-A8</td>
<td>192 GB</td>
<td>9</td>
<td>2 × Intel Xeon Gold 6148 + 2 × 3.84 TB NVMe SSDs</td>
<td>HPE Apollo 6500 Gen10</td>
</tr>
<tr>
<td>DB-s</td>
<td>128 GB</td>
<td>6</td>
<td>2 × Intel Xeon E5-2695 v3 + SSD</td>
<td>HPE ProLiant DL360</td>
</tr>
<tr>
<td>DB-h</td>
<td>128 GB</td>
<td>6</td>
<td>2 × Intel Xeon E5-2695 v3 + HDDs</td>
<td>HPE ProLiant DL380</td>
</tr>
<tr>
<td>Web</td>
<td>128 GB</td>
<td>6</td>
<td>2 × Intel Xeon E5-2695 v3</td>
<td>HPE ProLiant DL360</td>
</tr>
<tr>
<td>Othera</td>
<td>128 GB</td>
<td>16</td>
<td>2 × Intel Xeon E5-2695 v3</td>
<td>HPE ProLiant DL360, DL380</td>
</tr>
<tr>
<td>Gateway</td>
<td>64 GB</td>
<td>4</td>
<td>2 × Intel Xeon E5-2683 v3 (14c, 2.0/3.0 GHz, 35MB LLC)</td>
<td>HPE ProLiant DL380</td>
</tr>
<tr>
<td></td>
<td>64 GB</td>
<td>4</td>
<td>2 × Intel Xeon E5-2683 v3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>96 GB</td>
<td>2</td>
<td>2 × Intel Xeon</td>
<td></td>
</tr>
<tr>
<td>Storage</td>
<td>128 GB</td>
<td>5</td>
<td>2 × Intel Xeon E5-2680 v3 (12c, 2.5/3.3 GHz, 30 MB LLC)</td>
<td>Supermicro X10DRi</td>
</tr>
<tr>
<td></td>
<td>256 GB</td>
<td>15</td>
<td>2 × Intel Xeon E5-2680 v4 (14c, 2.4/3.3 GHz, 35 MB LLC)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>286.5 TB</td>
<td>920</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Other nodes = front end (2) + management/log (8) + boot (4) + MDS (4)
b. DDR4-2133
c. DDR4-2400
d. DDR4-2666
The Heart of Bridges-DL: NVIDIA Volta

New Streaming Multiprocessor (SM) architecture, introducing Tensor Cores, independent thread scheduling, combined L1 data cache and shared memory unit, and 50% higher energy efficiency over Pascal.

Tensor Cores accelerate deep learning training and inference, providing up to 12× and 6× higher peak flops respectively over the P100 GPUs currently available in XSEDE.

NVLink 2.0 delivering 300 GB/s total bandwidth per GV100, nearly 2× higher than P100.

HBM2 bandwidth and capacity increases: 900 GB/s and up to 32GB.

Enhanced Unified Memory and Address Translation Services improve accuracy of memory page migration by providing new access counters.

Cooperative Groups and New Cooperative Launch APIs expand the programming model to allow organizing groups of communicating threads.

Volta-Optimized Software includes new versions of frameworks and libraries optimized to take advantage of the Volta architecture: TensorFlow, Caffe2, MXNet, CNTK, cuDNN, cuBLAS, TensorRT, etc.

Training ResNet-50 with ImageNet:

V100 : 1075 images/s\textsuperscript{a}
P100 : 219 images/s\textsuperscript{b}
K80 : 52 images/s\textsuperscript{b}

\textsuperscript{a} https://devblogs.nvidia.com/tensor-core-ai-performance-milestones/
\textsuperscript{b} https://www.tensorflow.org/performance/benchmarks
Balancing AI Capability & Capacity: HPE Apollo 6500

**Bridges-DL adds 9 HPE Apollo 6500 Gen10 servers**

**Each HPE Apollo 6500 couples 8 NVIDIA Tesla V100 SXM2 GPUs**
- 40,960 CUDA cores and 5,120 tensor cores

**Performance:** 1 Pf/s mixed-precision tensor, 125 Tf/s 32b, 64 Tf/s 64b

**Memory:** 128 GB HBM2, 7.2 TB/s aggregate memory bandwidth

**2 × Intel Xeon Gold 6148 CPUs and 192 GB of DDR4-2666 RAM**
- 20c, 2.4–3.7 GHz, 27.5 MB L3, 3 UPI links

**2 × 4 TB NVMe SSDs** for user and system data

**1 × Intel Omni-Path host channel adapter**

**Hybrid cube-mesh topology** connecting the 8 V100 GPUs and 2 Xeon CPUs, using NVLink 2.0 between the GPUs and PCIe3 to the CPUs
**Maximum DL Capability: NVIDIA DGX-2**

**Couples 16 NVIDIA Tesla V100 SXM2 GPUs**
- 81,920 CUDA cores and 10,240 tensor cores

**Performance:** 2 Pf/s mixed-precision tensor, 251 Tf/s 32b, 125 Tf/s 64b

**Memory:** 512 GB HBM2, 14.4 TB/s aggregate memory bandwidth

**2 × Intel Xeon Platinum 8168 CPUs and 1.5 TB of DDR4-2666 RAM**
- 24c, 2.7–3.7 GHz, 33 MB L3, 3 UPI links

**2 × 960 GB NVMe SSDs** host the Ubuntu Linux OS

**8 × 3.84 TB NVMe SSDs** (aggregate ~30 TB) for user data

**8 × Mellanox ConnectX adapters** for EDR InfiniBand & 100 Gb/s Ethernet

The **NVSwitch** tightly couples the 16 V100 GPUs for capability & scaling
- Each of the 12 NVSwitch chips is an 18×18-port, fully-connected crossbar
- 50 GB/s/port and 900 GB/s/chip bidirectional bandwidths
- 2.4 TB/s system bisection bandwidth