Towards a Strategy for Data Sciences at UW

Albrecht Karle
Department of Physics
June 2017

- High performance computing infrastructure: Perspectives from Physics
- Existing infrastructure and projected future needs
- Other areas: computational astrophysics, quantum computing
- Cases for coordination:
  - Infrastructure
  - Data management
  - Machine learning
Computing in large physics “Big Science” projects: LHC experiments and IceCube

High Luminosity Large Hadron Collider

Experiments: CMS and ATLAS

IceCube at the South Pole
HL-LHC Challenge: Data!

- **Challenge**
  - Increasing data volume (x10)
  - Increasing event complexity
  - Increasing analysis sophistication
  - Multivariate analysis techniques require more compute resources
  - Moore’s law topping out
  - Budget limitations

- **R&D and Planning**
  - Beginning stages – opportunity
  - UW has tradition and expertise
  - Wisconsin should make tackling HLLHC computing challenge part of its research program
Raw Data Acquisition and Prompt Processing at Tier-0

Reconstruction takes place primarily at Tier-1

Analysis takes places primarily at Tier-2 and Tier-3

CMS Computing Systems

<table>
<thead>
<tr>
<th>Tier</th>
<th>US</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>T2</td>
<td>7</td>
<td>52</td>
</tr>
<tr>
<td>T3</td>
<td>30</td>
<td>63</td>
</tr>
</tbody>
</table>

Current Scale:
- 500K analysis jobs concurrently running
- 500PB storage in use worldwide

6/21/17
Sridhara Dasu (Wisconsin)
Wisconsin CMS Tier-2 Facility

✓ Compute (SL6 OS)
  • T2 HEP Pool – @11500 cores now
  • Added ~3100 cores last fall
  • Retired 400 oldest cores last fall

✓ Storage FS (Hadoop-2.0)
  • In use for 6+ years
  • Now @6.2 PB raw → 3.1 useable (2x replication)
  • Running in HA mode for past 5 months.

Needs to scale up by an order of magnitude for HL-LHC!
WIPAC (IceCube) Computing Facilities

UW has been lead institution for the IceCube construction and operation funded by largely NSF. UW’s WIPAC is the host institution and performs primary data processing and data management for the IceCube collaboration.

Overall infrastructure deployed across the various locations:
- GPU cluster (~450 GPUs) ~ 40% of power budget
- CPU cluster (~7600 cores) ~ 30% of power budget
- Disk storage (~7 PB) ~ 20% of power budget
- Infrastructure ~ 10% of power budget

Power/cooling/space are currently the limiting factor.

Distributed computing: a key component in IceCube data analysis

- About 50% of the cpu/gpu for IceCube is at UW-Madison, the rest is distributed worldwide.

- Acknowledge the advantage of having transparent access to OSG, XSEDE, etc using HTCondor related technologies.
WIPAC (IceCube) Computing Facilities

- **222 W Washington Av. datacenter**
  Core servers and data storage
  ~55 kW IT - at the limit of space and cooling

- **Chamberlin Hall** datacenters
  3rd Floor (3216)
  13 racks, ~115 kW IT capacity. Services: CPU/GPU cluster, data storage.
  4th Floor (4103)
  South Pole Test System <10 kW IT - inefficient room cooling

- **WID/MIR** - GZK9000 **GPU** cluster
  1st deployed in 2012, full upgrade in 2016, 1 rack, ~15 kW IT, shared with CHTC @ 30/70%

- **Near Future (6 months - 1 year):**
  starting design for new WIPAC datacenter space at CS 2360
  Goal: move the computing @ 222 into CS 2360, Aim for ~100 kW IT total at the new location.
Physics all: Computing requirements overview

Current power needs
   IceCube ~ 200 kW
   HEP ~ 160 kW
   Plasma ~ 5 kW
   Department ~ 2 kW

Total estimates with room for future expansion:

<table>
<thead>
<tr>
<th>PODs</th>
<th>Space (sqrft)</th>
<th>Power (kW)</th>
<th>Water (gpm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 - 8</td>
<td>3200 - 4300</td>
<td>900 - 1200</td>
<td>540 - 720</td>
</tr>
</tbody>
</table>

5-10 years forward view: larger share of compute could take place in the cloud, but Data I/O intensive workloads still have too high associated costs in the cloud Data storage & repositories still need to be hosted locally
Computational Astrophysics

Seen as an area of significant growth to enable new science:

- LSST (Large Synoptic Sky Survey), a very big new telescope with high data volume (15 TB/night)
  
  Keith Bechtol, faculty member starting next year

- Plasma astrophysics
Quantum Computing

Field of intense research and rapid development. Still in early stages it is based on alternate concepts of computing based on quantum mechanical phenomena.

Applications:

- Machine learning
- Big data
- Pattern recognition (principal component analysis)
- Drug discovery
- Cybersecurity/quantum internet/ factoring to break RSA encryption
- Quantum Chemistry
- Quantum materials simulation and development

Players:

- University & government labs
- Fortune 500 companies: IBM, Google, Intel, Microsoft, Lockheed,…
- Smaller companies/startups Dwave, IonQ, Rigetti, …

~ 5 M$/year in federally funded research:
- Quantum dots
- Superconductors
- Neutral atoms
Quantum Computing Investments

Government Investments:

- **Canada**: $140 M
- **USA**: $200+ M
- **UK**: $63 M
- **Netherlands**: $50 M
- **China**: $75 M
- **Singapore**: $170 M
- **Australia**: $40 M
Infrastructure

Physics is hosting some of the largest data processing/analysis infrastructures in campus.

We would benefit from participating in any data-infrastructure common infrastructure related discussions on campus.

- Datacenters
- Networking: roadmap for campus LAN/WAN evolution
- Long term archive services
- ...
**Data Science Areas**

**Data Management:** Effort to develop and maintain consistent data management plans
- IceCube total data output ~ 700 - 1000 TB per year
  - Need to host our own multi-petabyte repository.
- Leveraging large tape facilities at collaborating sites to provide the long term archive and preservation strategies (NERSC in the US and DESY-Zeuthen in Germany)
- Currently working in a data catalog project to ease discoverability of data products
- Would benefit from participating in a campus forum where similar experiences from other groups are shared, best practices, tools, etc

**Machine Learning**
- A growing area in IceCube - larger number of analysis exploring the capabilities of new deep neural networks tools for event selection and reconstruction.
- Having access to a focused multidisciplinary expert group in campus could be beneficial.
Conclusion

Possible cases for coordination towards a strategy in data sciences:

- Infrastructure
- Data management
- Machine learning
- ...