Assessing the science requirements for Earth system science in the NWSC-3 Procurement

Dr. Richard Loft
Director of Technology Development
Computational and Information Systems Laboratory
National Center for Atmospheric Research

November 13, 2018
IEEE Supercomputing, Dallas Texas
• Background
• Emerging Applications
• Workload Study
• Strawman Design
• Q&A
Cheyenne physical infrastructure

<table>
<thead>
<tr>
<th>Resource</th>
<th># Racks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cheyenne</td>
<td>14 HPE 8600 E-Cells each containing 2 water-cooled E-Racks &amp; heat exchanger, and 16 Mellanox 36-port EDR InfiniBand switches&lt;br&gt;2 air-cooled storage &amp; service racks including login nodes</td>
</tr>
<tr>
<td>Casper</td>
<td>3 PCPC racks; 100 GigE &amp; FDR interconnect</td>
</tr>
<tr>
<td>GLADE</td>
<td>8 DDN SFA14KXe racks containing 32 NSD servers and storage</td>
</tr>
</tbody>
</table>

Total Power

- **HPC**: 1.75 MW
- **Casper**: ~0.11 MW
- **GLADE**: 0.21 MW

Total Power ~2.0 MW
Cheyenne Environment

R&E Networks

Casper (DAV)
- 24 nodes (Various)

Cheyenne (HPC)
- 5.4 petaflops
- 147k Xeon v4 cores

EDR fabric

40/100 Gbps Ethernet

Data Movers

Cold Cache (Tape/Disk)

~100 PB Tape Archive

Campbell Storage
- 20 PB

Hot Cache (NVMe)
- 0.45 PB SSD

Warm Cache (HPC Disk)
- 38 PB

GLADE
NWSC Science Requirements Analysis

• Workload Study
  – Some data already collected

• Technical input
  – Workload study
  – Scientific research objectives from individual PI’s
  – User surveys

• Science Requirements Advisory Panel will
  – Be comprised of internal and external members from community
  – Collect and evaluate technical input
  – Advise NWSC-3 RFP development and procurement decisions
Outline

• Background
• Emerging Applications
• Workload Study
• Strawman Design
• Q&A
NCAR future research priorities

- Air quality prediction
- Extreme events in climate
- Coupled prediction of the Arctic
- Space Weather Predictability
- Tropical cyclone predictability
What’s driving NCAR future applications? ESP!

• **Then**
  – Weather prediction (5-10 days)
  – GAP
  – Climate projections (decades-centuries)

• **Division between meteorology and climate is breaking down**
  – Discoveries in the ocean

• **Now: earth system prediction (ESP) filling that GAP**
  – Subseasonal (Weeks)
  – Seasonal (Months)
  – Climate predictions (years to decadal)
Computing implications of NCAR’s ESP goals

• Coupled models of atmosphere, ocean, sea ice, geospace will become the norm.
• Data assimilation will become much more prevalent.
• Sweet spot of workflow of global models needs to drop from 100 km to 25 km.
• Regionally refined resolutions will need to drop to (1-3 km).
• Goals represent a weak scaling pivot to a much higher level of parallelism for our Earth system models.
• Data assimilation will require innovative job scheduling and data-intensive features.
Emerging Applications

• **Machine Learning (ML > DL)**
  – Data pre-processing (cleaning up observations)
  – Modeling (physics emulation)
  – Data post processing (compression, feature detection)
  – Fault detection (smart CI systems)

• **GPU computing**
  – Whole models (MHD, MPAS-Atmosphere,...)
  – Compute-intensive components (chemistry, LES models)

• **Persistent Data Analytics**
  – Scalable data storage format (data cubes)
  – Parallel workflow orchestration (Dask)
  – ES science analytics tool chain (NCL)

• **High throughput computing**
  – Data assimilation system development
  – Build/Test
Outline

- Background
- Emerging Applications
- Workload Study
- Strawman Design
- Q&A
Cheyenne usage reflects its mission to serve the atmospheric sciences.

Climate: 50% of NCAR’s workload.
Besides CESM, 30+ other applications and models identified across 177 projects.

These 177 projects represent 93% of Cheyenne’s reported use.
A project’s depth is defined as the size of its largest job in nodes (rounded up to the next greater power of 2). Note: Yellowstone nodes had 16 cores.

64% of projects never need more than 64 nodes. These projects use only 13% of the total node-hours.

33% of projects use a max of 65-1024 nodes. These projects use 71% of the total node-hours.

3% of projects use a max of more than 1024 nodes. These projects use 16% of the total node-hours.
A project’s depth is defined as the size of its largest job (rounded to the next greater power of 2).

Note the similarity in usage by projects compared to Yellowstone use.
### 100 km CESM on Cheyenne: greater capability

<table>
<thead>
<tr>
<th>NCAR System</th>
<th>Intel Xeon Processor</th>
<th>CESM Version</th>
<th>Capability (sim yr/day)</th>
<th>Cost (core-hr per sim yr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cheyenne</td>
<td>18c (v4)</td>
<td>CESM2</td>
<td>30</td>
<td>3500</td>
</tr>
<tr>
<td>Yellowstone</td>
<td>8c (v2)</td>
<td>CESM2</td>
<td>19.6</td>
<td>5167</td>
</tr>
<tr>
<td>Yellowstone</td>
<td>8c (v2)</td>
<td>CESM1</td>
<td>10.6</td>
<td>1521</td>
</tr>
</tbody>
</table>

### 100 km CESM take-aways:

**CESM is 48% more efficient on Cheyenne compared to Yellowstone.**

**CESM-2 on Cheyenne can deliver 2.8x the capability, compared to CESM1 on Yellowstone.**
Cheyenne workload oozing to higher parallelism

Job size in Cheyenne node equivalents (CN = 2.25x YS)
Cheyenne job data show a pronounced capacity use case: 81% of jobs are short, 1-node and consume 3% of total core-hours.
GLADE storage growth over time

CH (/scratch+ /proj) = 0.42 PB/month

YS (/scratch+/proj) = 0.27 PB/month

CH /glade/scratch

YS /glade/scratch
(YS) Yellowstone -> Cheyenne (CH) data: what can we conclude about future HPC storage demands?

- A **3x** increase in *estimated* workload capacity resulted in an *apparent* increase of
  - **2x** in /scratch asymptotic size
  - **1.8x** in /scratch + /project storage growth rate
- Linear/power law projections suggest that a 3x increase in workload capacity for NWSC3 would equate to:
  - /scratch asymptote should be **~20-25 petabytes**
  - An expected growth rate of **0.76 - 0.87 PB/month** in /scratch + /project storage
Outline

- Background
- Emerging Applications
- Workload Study
- Strawman Design
- Q&A
NWSC-3 Design Strawman (History + Trends)

- R&E Networks
- Commercial Cloud
- Off Prem Cloud
- Ethemet
- Data Movers
- Campaign Storage
- Cold Cache (Tape/Disk) Capacity
- Campaign
- HPC System
- CPU
- GPU
- HPC fabric
- Warm Cache (HPC Disk)
- Hot Cache (NVMe)
- Tape-Archive
- On Prem Cloud
- Cloud R&E Networks
- Commercial Cloud
- Off Prem Cloud
- Ethernet
- Data Movers
- Campaign Storage
- Cold Cache (Tape/Disk) Capacity
- Campaign
- HPC System
- CPU
- GPU
- HPC fabric
- Warm Cache (HPC Disk)
- Hot Cache (NVMe)
Outline

• Background
• Emerging Applications
• Workload Study
• Strawman Design
• Q&A
Breakdown of climate work on Cheyenne

- CESM
- CESM variant
- WACCM
- WACCM-X
- POP2
- CAM & variants
- CLM