Cheyenne

NCAR’s Next-Generation Data-Centric Supercomputing Environment

David Hart, NCAR/CISL User Services Manager
June 23, 2016
History of computing at NCAR

Red text indicates those systems that are currently in operation.
Cheyenne
Planned production, January 2017 – 2021

• **Scientific Computation Nodes**
  – SGI ICE XA cluster
  – 4,032 dual-socket nodes
  – 18-core, 2.3-GHz Intel Xeon E5-2697v4 processors
  – 145,152 “Broadwell” cores total
  – 5.34 PFLOPs peak – 1.325 TFLOPs per node!
  – 313 TB total memory (3,164 64-GB and 864 128-GB nodes)
  – >2.45 Yellowstone equivalents

• **High-Performance Interconnect**
  – Mellanox EDR InfiniBand
  – 9-D enhanced hypercube topology
  – 100-Gbps link bandwidth — 0.5 µs latency
  – 36 TB/s bisection bandwidth
  – 224 36-port switches, no director switches

• **Login Nodes**
  – 6 nodes with 2.3-GHz Intel Xeon E5-2697v4 processors
  – 36 cores & 256 GB memory per node

• **Service Nodes (scheduler, license servers)**
  – 4 nodes; Intel Broadwell processors
  – 36 cores & 64 GB memory per node
Cheyenne compute nodes

Cheyenne represents 1.2 billion core-hours per year!

• 4,032 dual-socket nodes
• Intel E5-2697v4 processors
  – 2.3-GHz, 18-core “Broadwell” processors
• Four logical 2-socket nodes per blade
  – SGI ICE-XA “Hawkins” blades
• Memory
  – Four DDR4-2400 DIMMs per socket
  – 864 nodes with 128 GB memory
    • 8x 16-GB (dual-ranked)
  – 3,168 nodes with 64 GB memory
    • 8x 8-GB (single-ranked)
• Interconnect
  – Four HCA Mezzanine Card Slots
  – Mellanox ConnectX-4 EDR
  – 9D enhanced hypercube, a scalable topology well-suited to large node-count MPI jobs

Cheyenne Building Blocks

• 1 socket has 18 cores
• 1 compute node has 2 sockets
• 1 blade has 4 compute nodes
• 1 blade enclosure has 9 blades
• 1 E-Rack has 4 enclosures
• 1 E-Cell has 2 E-Racks
• Cheyenne has 14 E-Cells
Cheyenne software

- **Compilers, Libraries, Debugger & Performance Tools**
  - Intel Parallel Studio XE Cluster
    - Fortran, C++, performance & MPI libraries, trace collector & analyzer
  - Intel VTune Amplifier XE performance optimizer
  - PGI CDK (Fortran, C, C++, pgdbg debugger, pgprof)
  - Allinea Forge and Performance Reports
  - SGI Message Passing Toolkit (MPT)

- **System Software**
  - Altair PBS Pro Batch Subsystem / Resource Manager
  - SuSE Linux (Operating System)
  - IBM Spectrum Scale Parallel File System software (GPFS)
  - Mellanox Unified Fabric Manager
  - SGI Management Center (Cluster administration)
  - SGI Foundation Software (tools/utilities)
NWSC-2 test systems

Laramie
- SGI ICE XA (air cooled)
- 72 dual-socket nodes
- 18-core, 2.1-GHz Intel E5-2695v4
- 2,592 cores total
- 64 GB memory per node
- Mellanox EDR InfiniBand
  - 3D enhanced hypercube
- 87 TFLOPS peak
  - 15 TFLOPS greater than Bluefire!

Picnic+
- 1x DDN SFA14KX
- 1x 84-slot drive chassis
- 40x 8-TB NL SAS drives
- 2x 400-GB SSDs (metadata)
- 2x NSD servers
- 250 TB usable capacity
  - 100 TB greater than Bluefire!
Cheyenne deployment timeline

- Facility and networking infrastructure upgrades at NWSC started early 2016.
  - Will include 100-GbE links to the outside world
- Test system, Laramie, arrives in July.
- Cheyenne hardware to be assembled in Chippewa Falls and to undergo factory testing in August.
  - Will include HPL run for fall Top500 list
- Cheyenne hardware to be shipped to NWSC in September.
- From September through December, system to go through installation, integration with GLADE, and acceptance testing (including 30-day availability testing).
- Targeting NCAR acceptance by mid-December
- Start of production: January 2017
  - Yellowstone continues in production through December 2017.
- City of Cheyenne celebrates its 150th anniversary in 2017.
NWSC-2 file system resource

- **21-PB DDN Storage Fusion Architecture (SFA) system**
  - 4x SFA14KX units
  - 4x20 84-slot drive chassis
  - 3,360x 8-TB NL SAS drives (2x expandable)
  - 26.8 PB raw capacity

- **200 GB/s aggregate I/O bandwidth**
  - 48x 800-GB, mixed-use SSDs for metadata
  - 24x NSD servers (Dell PowerEdge R630)
  - EDR InfiniBand and 40-Gig Ethernet connections

- **Total integrated GLADE capacity: 37 PB**
  - Integrates with existing 16-PB file system
  - Expandable to 42-PB (58-PB total) by adding drives

- **IBM Spectrum Scale software**
  - Formerly GPFS
  - RedHat Enterprise Linux OS
GLADE-HPC integration

• NSD servers are all connected to 40-GbE network
  – Will route traffic between the two IB networks
  – Data traffic will be routed between the old and new NSD servers until Yellowstone retires
• File service to gateways, RDA, ESG and HPSS over 40 GbE
• EDR InfiniBand network
  – Connect 24 new NSD servers directly to Cheyenne’s EDR fabric
  – Connect 20 old NSD servers directly to Cheyenne fabric once Yellowstone is decommissioned (2018)
• Add to configuration
  – 3 new utility nodes to handle GPFS cluster management
• HPSS access through the gladedm nodes
  – 4 new gladedm nodes for use with Cheyenne
GLADE file system tiers

GLADE GPFS cluster

$SCRATCH
DDN SFA14KX
15 PB

$PROJECT
DDN SFA14KX
5 PB

$HOME
$APPS
DDN
100 TB

DCS3700
16 PB
NCAR disk storage capacity profile

Total Centralized File system Storage

- NWSC-2 Options
- GLADE (NWSC-2)
- GLADE (NWSC-1)
- GLADE (ML)

- NWSC-2 option to double new storage capacity +21 PB
- November 2016

- GLADE (NWSC-2) +21 PB + 200 GB/s via EDR InfiniBand
- GLADE (NWSC-1) + 16.4 PB + 90 GB/s via FDR InfiniBand

- GLADE (Mesa Lab) 6 GB/s

Usable capacity (PB)

Jan-10 Jan-11 Jan-12 Jan-13 Jan-14 Jan-15 Jan-16 Jan-17 Jan-18 Jan-19 Jan-20 Jan-21

Usable capacity profile:
- 58 PB
- 37 PB
Data analysis & visualization resources

- **Geyser: Large Memory System**
  - 16 IBM x3850 X5 nodes
  - Intel E7-4870 (Westmere-EX) processors
  - 40 2.4-GHz cores, 1 TB memory per node
  - 1 NVIDIA K5000 GPU per node
  - Mellanox FDR InfiniBand interconnect

- **Caldera: GPGPU & Visualization System**
  - 30 IBM dx360 M4 nodes
  - Intel Xeon E5-2670 (Sandy Bridge) processors
  - 16 2.6-GHz cores, 64 GB memory per node
  - 2 NVIDIA Tesla K20X GPUs on 16 nodes
  - Mellanox FDR InfiniBand interconnect
NCAR HPSS archive resource

• **NWSC**
  - Two SL8500 robotic libraries
  - 20,000 cartridge capacity
  - 46 T10000C tape drives — 240 MB/s I/O
    • 5-TB “C” cartridges
  - 46 T10000D tape drives — 240 MB/s I/O
    • 8-TB “D” cartridges
  - 160 PB capacity
  - Current total holdings: **60 PB**
  - Current growth rate: **~12 PB/year**

• **Mesa Lab**
  - Two SL8500 robotic libraries
  - 15,000 cartridge capacity
  - 15-PB capacity for disaster recovery data

• **Upgrade planned for late 2017**
HPSS archive stored data profile

60 PB reached as of May 8, 2016

40 PB added since start of Yellowstone
NCAR data-intensive supercomputing environment

GLADE
Central disk resource
36.4 PB, 90/200 GB/s GPFS

Cheyenne
5.34 PFLOPS peak

Yellowstone
1.5 PFLOPS peak

EDR / FDR InfiniBand and 40-Gb Ethernet

High-bandwidth Low-latency HPC and I/O Networks

RDA, Climate Data Service

Data Transfer Services
40 GB/s

HPSS Archive
175-190 PB capacity
60 PB stored, 12.5 GB/s
~14 PB/yr growth

Remote Vis
Partner Sites
XSEDE Sites
NWSC-2 maintenance & support

- **Contractual reliability guarantees**
  - HPC hardware and software to exceed 98% availability and >384 hours (16 days) Mean Time Between System Interrupt
  - CFDS resources hardware and software to exceed 99% Availability and >576 hours (24 days) Mean Time Between System Interrupt
  - Anticipate equipment unit availability and MTBF to well exceed 99.9% and 1000s of hours

- **Hardware maintenance**
  - 24x7, four-hour, on-site hardware maintenance for all components
  - Eaton provides support for CDU (cooling units)
# Cheyenne physical infrastructure

<table>
<thead>
<tr>
<th>Resource</th>
<th># Racks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cheyenne</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>SGI ICE XA E-Cells each containing 2 water-cooled E-Racks &amp; heat exchanger, and 16 Mellanox 36-port EDR InfiniBand switches</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>air-cooled storage &amp; service racks including login nodes</td>
</tr>
<tr>
<td>GLADE+</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>DDN SFA14KX racks containing 24 NSD servers and storage</td>
</tr>
<tr>
<td>GLADE</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>NSD server, controller, and storage racks</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>19” Rack (I/O aggregator nodes, management, switches)</td>
</tr>
<tr>
<td>DAV</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>IBM iDataPlex Rack (Caldera, Pronghorn)</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>19” Racks (Geyser, management, IB switch)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total Power</th>
<th>~2.0 MW</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPC</td>
<td>1.75 MW</td>
</tr>
<tr>
<td>GLADE</td>
<td>0.21 MW</td>
</tr>
<tr>
<td>DAV</td>
<td>0.063 MW</td>
</tr>
</tbody>
</table>
### Cheyenne power efficiency

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Processor</td>
<td>Xeon E5-2697v4 2.3 GHz</td>
<td>Xeon E5-2670 2.6 GHz</td>
<td>POWER6 4.7 GHz</td>
</tr>
<tr>
<td>Total Batch Processor Cores</td>
<td>145,152</td>
<td>72,288</td>
<td>3,744</td>
</tr>
<tr>
<td>HPC portion peak PFLOPs</td>
<td>5.342</td>
<td>1.510</td>
<td>0.072</td>
</tr>
<tr>
<td>Power consumption</td>
<td>1.75 MW</td>
<td>1.4 MW</td>
<td>0.54 MW</td>
</tr>
<tr>
<td>Watts/peak GFLOPS</td>
<td>0.33</td>
<td>0.93</td>
<td>7.5</td>
</tr>
<tr>
<td>Peak GFLOPS/watt</td>
<td>3.05</td>
<td>1.08</td>
<td>0.133</td>
</tr>
<tr>
<td>Average workload floating point efficiency</td>
<td>1.1% (estimate)</td>
<td>1.56% (measured)</td>
<td>3.9% (measured)</td>
</tr>
<tr>
<td>Sustained MFLOPS/watt (on NCAR workload)</td>
<td>~34</td>
<td>~17</td>
<td>5.2</td>
</tr>
<tr>
<td>Bluefire-equivalents</td>
<td>70.8</td>
<td>28.9</td>
<td>1</td>
</tr>
<tr>
<td>Yellowstone-equivalents</td>
<td>&gt;2.45</td>
<td>1</td>
<td>0.035</td>
</tr>
</tbody>
</table>

For 3.2x more power, Cheyenne will deliver 71x more computational performance than Bluefire.
Historical power efficiency on NCAR workload

- Cray J90
- SGI Origin2000
- Compaq ES40
- SGI Origin3800
- IBM POWER4 p690
- IBM AMD/Opteron Linux
- IBM BlueGene/L (frost)
- IBM POWER5 p575 (bluevista)
- IBM POWER5+ p575 (blueice)

- Yellowstone ~20 MFLOPs/W
- Lynx ~14 MFLOPs/W
- Frost ~5.5 MFLOPs/W
- Bluefire ~7.4 MFLOPs/W
- Bluevista / Blueice ~1.4 MFLOPs/W

Cheyenne ~34 MFLOPs/W
User communities for CISL HPC

NCAR supports four user communities through policies established via various agreements with NSF or approved by NSF.

- **University research**
  - Eligibility: “In general, any U.S.-based researcher with an NSF award in the atmospheric, geospace or closely related sciences is eligible to apply for a University Community allocation.”
  - Large requests reviewed by CISL HPC Allocation Panel (CHAP)
  - Small allocations for projects with NSF awards usually processed in ~1 day
  - Small allocations, without NSF award, also available to graduate students, post-docs, or new faculty; also, instructional allocations (classroom, tutorial) in eligible field.

- **Climate Simulation Laboratory**
  - Supports large-scale, long-running climate simulations
  - Eligibility otherwise similar to University allocations
  - Also supports large annual allocation to CESM Community
  - Requests reviewed by CHAP

- **NCAR Lab and NCAR Strategic Capability (NSC) activities**
  - NCAR staff may engage collaborators from outside of NCAR
  - Large requests reviewed by internal NCAR Allocation Review Panel, approved by NCAR Executive Committee

- **Wyoming-NCAR Alliance**
  - Must be led by U Wyoming researcher
  - Must be in the Geosciences or related fields (including solid Earth geophysics)
  - Any source of funding support
  - Large requests reviewed by Wyoming Resource Allocation Panel (WRAP)
HPC allocations

The percentages available to each of the four user communities remain roughly the same across both Yellowstone and Cheyenne.

With Cheyenne, the Antarctic Mesoscale Prediction System (AMPS) will no longer have a separate cluster and will run within the production workload, using Laramie as a backup, if needed.
HPC allocations and use in 2017

• In 2017, both Yellowstone and Cheyenne will be in production.

• In 2018, Yellowstone decommissioning represents a 30% drop in compute capacity.

• Allocations and access to be managed to target overlap year capacity to fixed-duration activities.
  – All while maintaining target usage percentages to all stakeholder communities!
Allocations in late 2016

• All communities have fall allocations opportunities—Wyoming, University, NCAR, and Climate Simulation Lab (CSL).

• Most fall allocations will be made on Cheyenne system.
  – Current, active Yellowstone projects may be able to request modest amounts of “wrap-up” time.
  – Rest of Yellowstone to support existing allocations.

• Final approach may vary depending on various factors—e.g., unspent allocations, desire by groups to move work to Cheyenne, and so on.
Accelerated Scientific Discovery

- ASD will provide a small number of “shovel-ready” projects that can consume lots of Cheyenne in a short period.
- Tentatively, 10–12 projects and ~250 million core-hours (total) in two to three months.
  - Split between university and NCAR projects.
  - So average of ~20M core-hours per project.
- University projects to be reviewed by CHAP
- NCAR projects to be reviewed by NSC panel
# Tentative ASD schedule

<table>
<thead>
<tr>
<th>DATE</th>
<th>DEADLINE/MILESTONE</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUG 29</td>
<td>University ASD submission deadline; NSC/NCAR ASD submission deadline</td>
</tr>
<tr>
<td>SEP 12</td>
<td>CHAP submission deadline</td>
</tr>
<tr>
<td>~ OCT 7</td>
<td>University &amp; NCAR ASD recommendations</td>
</tr>
<tr>
<td>NOV 1</td>
<td>ASD teams start on Laramie test system</td>
</tr>
<tr>
<td>DEC 1</td>
<td>ASD teams ready for Cheyenne</td>
</tr>
<tr>
<td>DEC 18</td>
<td>Cheyenne acceptance (official plan) ASD users get access</td>
</tr>
<tr>
<td>JAN 1, 2017</td>
<td>Cheyenne enters production ASD projects start</td>
</tr>
</tbody>
</table>
## Cheyenne v. Yellowstone

To estimate the number of core-hours needed on Cheyenne, multiply Yellowstone core-hours by 0.82.

\[ 20M \text{ Yellowstone core-hours} \times 0.82 = 16.4M \text{ Cheyenne core-hours} \]

<table>
<thead>
<tr>
<th></th>
<th>CHEYENNE</th>
<th>YELLOSTONE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total cores (batch nodes)</td>
<td>145,152 (2.01x)</td>
<td>72,256</td>
</tr>
<tr>
<td>Total memory (batch nodes)</td>
<td>313 TB (2.16x)</td>
<td>145 TB</td>
</tr>
<tr>
<td>Peak PFLOPS</td>
<td>5.34</td>
<td>1.5</td>
</tr>
<tr>
<td>Batch node racks</td>
<td>28 (0.44x)</td>
<td>63</td>
</tr>
<tr>
<td>InfiniBand generation</td>
<td>EDR (100 Gbps)</td>
<td>FDR (56 Gbps)</td>
</tr>
<tr>
<td>Interconnect topology</td>
<td>9D enhanced hypercube</td>
<td>Full fat tree</td>
</tr>
<tr>
<td>IB fabric director switches</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>HPC power required</td>
<td>1.75 MW</td>
<td>1.4 MW</td>
</tr>
<tr>
<td>Yellowstone-equivalents</td>
<td>2.45</td>
<td>1</td>
</tr>
<tr>
<td>Core-hours : YS core-hour</td>
<td>0.82</td>
<td>1</td>
</tr>
</tbody>
</table>
CISL HPC and storage roadmap

- **Cheyenne (NWSC-2)**
  - GLADE-1 (16 PB)
  - GLADE-2a (21 PB)
  - GLADE-2b (+21 PB)
  - Many-core System
  - DAV (for NWSC-2) with NVM
  - Procurement
  - RFI

- **Yellowstone (NWSC-1)**
  - GLADE-1 (reduced QoS)
  - Procurement for tape archive
  - HPSS w/ “D” Tapes and Drives

- **NWSC-2 Phase 2**
  - Procurement
  - Sys and Facility Prep

- **Storage**
  - GLADE-2c

- **HPC**
  - CISL HPC and storage roadmap

Timeline:
- 1/17
- 6/17
- 1/18
- 6/18
- 1/19
- 6/19
- 1/20
- 6/20
- 1/21
- 6/21
QUESTIONS?