Early Experiences with Yellowstone

John Dennis
Outline

• ASD project

• Why does my code sometimes run slowly on Yellowstone?
  – Infiniband routing

• Who stole my Supercomputer Performance?
  – POPperf benchmark
  – OS jitter
Advanced Scientific Discovery project

- Opportunity to use a large piece of newly installed Yellowstone
- “Meso- to planetary-scale processes in a global ultra-high resolution model”
- R. Justin Small, Bailey, Bryan, Danabasoglu, Holland, Jochum, Lawrence, Park, Peacock, Tomas, Tribbia, Dennis (NCAR), Saravanan (Texas A&M), Schneider (Hawaii), Kwon (WHOI)
- 47.1 M core hours
  - 25.2 M (2 months) [tuning]
  - 21.9 M (15 months)
High-resolution ASD simulation

- ~80 years complete
- CAM5-SE (atmosphere model)
  - 28km resolution
  - Scalable spectral-element dynamical core
  - CAM5 physics
  - Fully prognostic aerosol (~50 tracers)
- POP (ocean model)
  - 11km resolution
  - 62 vertical levels
- CICE (sea-ice model)
  - 11km resolution
- CLM (land model)
  - 28km resolution
Computational Aspects of ASD simulation

- **General statistics:**
  - 2.0 simulated years per day (SYPD)
  - 23,404 cores
  - 1 TB of data generated per day

- **Component configuration**
  - 11km Ocean model (6,124 cores)
  - 11km Sea-ice model (16,295 cores)
  - 28km Atmosphere (17,280 cores)
  - 28km Land (900 cores)
  - Coupler (10,800 cores)
Execution time for ASD simulation

- CAM (17280x1) 52%
- POP (6124x1) 25%
- CLM (900x1)
- CICE (16295x1) 11%
- CPL (10800x1)

# of cores
Execution time for ASD on Yellowstone

Large variability in execution time

Graph showing execution time for non-I/O CESM day with dates ranging from 11/25 to 01/06.
Write bandwidth for ASD simulation on Yellowstone

6.4% overhead

21% overhead
Why did we not use more of Yellowstone?

- **System stability**
  - Significant performance loss (8x) in MPI_Allreduce on occasion (~1%)
  - Collective communication offload failures (33%)

- **Queue access for larger than 22k cores**

- **OS jitter sensitivity** [CAM-SE, POP, CICE]

- **Infiniband routing table imbalances**

- **I/O overhead** (6.4%)
Outline

• ASD project
• Why does my code sometimes run slowly on Yellowstone?
  – Infiniband routing
• Who stole my Supercomputer Performance?
  – POPperf benchmark
  – OS jitter
Execution time for ASD on Yellowstone

Large variability in execution time
[Our] global 3-km simulations that we'd like to run on 16384 cores, ... these simulations have been consistently running about 60% slower than in the past.

-Michael Duda (April 24, 2013) -
Top of the Rack (TOR) switch in Yellowstone

Which upward links for packets with destination: A1?

Compute nodes

TOR: ys58ib1

Other switches

Up the tree (upward)

Down the tree (downward)
Load-imbalance in routing table

Unbalanced upward links
Balanced upward links

Downward

Upward

Change in the routing table imbalance

High-performance computing Advisory Panel (CHAP)
Impact of Load-imbalance in routing tables

• Contention for network bandwidth
  – Negative impact on bisection bandwidth limited codes

• Nature of contention changes over time
  – Routing reconfigured every time a hardware failure or error occurs
  – Has occurred at least 61 times in 53 days
    ~ 20.8 hours between reconfiguration
  – Cost of routing reconfiguration?
Outline

• ASD project
• Why does my code sometimes run slowly on Yellowstone?
  – Infiniband routing
• Who stole my Supercomputer Performance?
  – POPperf benchmark
  – OS jitter
Parallel Ocean Program (POP)

- Community Earth System Model (CESM) ocean model
- High-resolution 0.1°
  - Separation between grid-points ~ 10 km
  - 3600x2400x42 grid-points
  - Eddy-resolving
- Computationally challenging (barotropic)
  - Latency sensitive
  - OS-jitter sensitive
**POPperf benchmark**

**Simulation rate**

Simulation rate for POPperf 0.1 deg benchmark

- Janus [Dell w/QDR]
- Hopper [Cray XE6]
- Pleiades [SGI w/QDR]
- Yellowstone [IBM w/FDR]
- Yellowstone w/FCA [IBM w/FDR]
**POPperf benchmark**

*(barotropic section)*

Barotropic execution time for POPperf 0.1 deg benchmark

- **Janus [Dell w/QDR]**
- **Hopper [Cray XE6]**
- **Pleiades [SGI w/QDR]**
- **Yellowstone [IBM w/FDR]**
- **Yellowstone w/FCA [IBM w/FDR]**

**Offload of MPI reduction**

*Graph showing execution time for different number of compute cores.***
Outline

• ASD project
• Why does my code sometimes run slowly on Yellowstone?
  – Infiniband routing
• Who stole my Supercomputer Performance?
  – POPperf benchmark
  – OS jitter
Tracing HOMME on Yellowstone

- Utilizing Extrae tracing
  - Collaboration with Barcelona Super Computer Center (Labrata, Gimenez)
  - Very detailed analysis
  - “Performance analysis using a microscope”

- HOMME configuration (1080 cores)
  - Similar weak scaled configuration to the atmosphere model running under ASD project
  - 5 elements per core (ne=30)
Computation time analysis

Useful duration (user level code) & Histogram

395; 580 us

100 - 350us

Imbalanced in completion time!
Can we detect the noise?

MPI call duration

Preemptions @ computation

Preemptions @ MPI

OS-jitter:

Frequency: 0.08 sec
Duration: 400 usec
Eliminating noise at computations

- Use instructions as computation “cleaned” reference – approximation (assumes constant Instructions per cycle)
Conclusions

• Yellowstone has enabled very large scale simulation capability at NCAR
  – Much better experience then other large scale computational campaign

• Yellowstone still has issues
  – Routing table imbalances
  – Frequent reconfiguration of routing table
  – OS jitter
**POPperf benchmark (3D-update section)**

3D-update execution time for POPperf 0.1 deg benchmark

- **Janus [Dell w/QDR]**
- **Hopper [Cray XE6]**
- **Pleiades [SGI w/QDR]**
- **Yellowstone [IBM w/FDR]**
- **Yellowstone w/FCA [IBM w/FDR]**

**Graph Details:**
- **Y-axis:** Seconds per day
- **X-axis:** Number of compute cores (128, 256, 512, 1K, 2K, 4K, 8K, 16K, 32K)

The graph compares the performance of different computing systems in terms of 3D-update execution time, with each system represented by a distinct line.
High-resolution ASD simulation

- Improved mean-climate then previous HR runs
- Improved Arctic sea-ice extent
- Ocean temperature bias in the Southern ocean ACC
Picture of OS-jitter on Yellowstone (Extrae + Paraver) [J. Labarta]

400 microsec OS-jitter
## Computational characteristics of 0.1° POP

<table>
<thead>
<tr>
<th>Section of code</th>
<th>Description</th>
<th>At large core count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baroclinic</td>
<td>Floating-point intensive</td>
<td>Mostly floating-point dominated</td>
</tr>
<tr>
<td>Barotropic (*)</td>
<td>2-dimensional CG solver</td>
<td>Very latency sensitive + MPI reduction</td>
</tr>
<tr>
<td>3D-update (*)</td>
<td>Nearest neighbor communication pattern</td>
<td>Somewhat latency sensitive</td>
</tr>
</tbody>
</table>

(*) Key to scaling to very large core counts