TEXAS ADVANCED COMPUTING CENTER
WWW.TACC.UTEXAS.EDU

TEXAS
The University of Texas at Austin

Impact of Meltdown/Spectre Patches on HPC Application Performance

IXPUG Spring Conference
Cineca

PRESENTED BY:
John Cazes
cazes@tacc.utexas.edu

Meltdown Exploit

Exploits out-of-order execution to load protected memory into data cache without checking permissions

Lipp, M., Schwarz, M., Gruss, D., Prescher, .,Haas, W., Mangard, S., Kocher, P., Genkin, D.,Yarom, Y.&Hamburg, M. Meltdown. ArXiv eprints. arXiv: 1801.01207 (Jan. 2018).

Spectre Exploit Variant One

Uses speculative execution techniques to leak privileged information

Kocher, P., Genkin, D., Gruss, D., Haas, W., Hamburg,M., Lipp, M., Mangard, S., Prescher, T.,Schwarz, M.&Yarom, Y. Spectre Attacks: ExploitingSpeculative Execution. ArXiv e-prints. arXiv: 1801.01203 (Jan. 2018).

Meltdown Kernel Patch

Implement Kernel Page Table Isolation

- System calls cause a “real” context switch
- Caches are flushed
- Can be disabled

Spectre Variant 1 Kernel Patch

Implement kernel “load fences”

- Prevents out-of-bound loads into cache
- Always enabled

https://www.redhat.com/en/blog/what-are-meltdown-and-spectre-heres-what-you-need-know

Spectre Variant 2 Kernel Patches

Indirect Branch Restricted Speculation

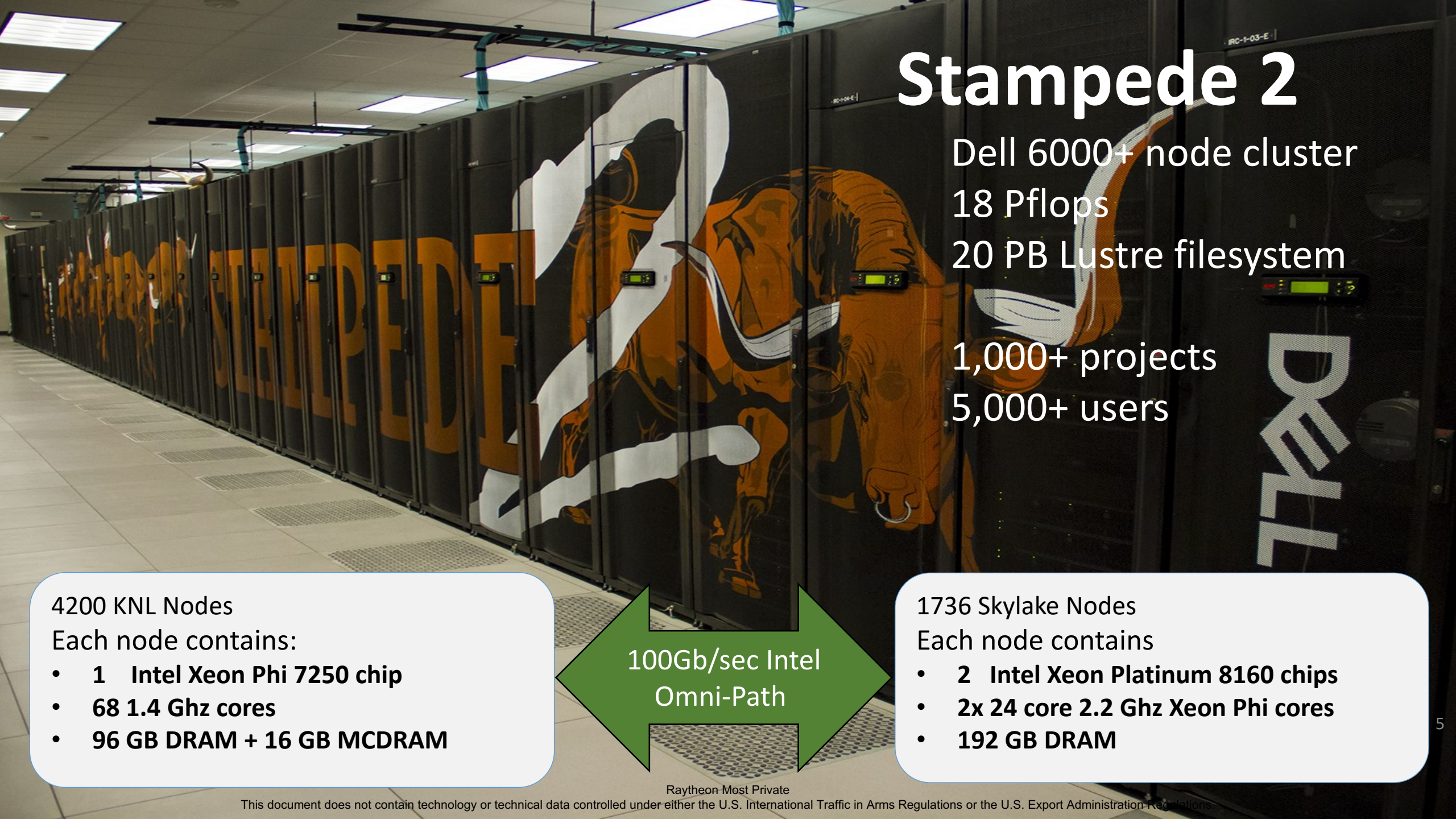
- Restricts kernel space speculative execution
- Can restrict user space speculative execution
- Can be disabled

Indirect Branch Prediction Barriers

- Flushes results of speculative executions if there is a context switch
- Can be disabled

Requires a microcode update – not available yet for SKX/KNL

https://access.redhat.com/articles/3311301



Stampede 2

Dell 6000+ node cluster
18 Pflops
20 PB Lustre filesystem

1,000+ projects
5,000+ users

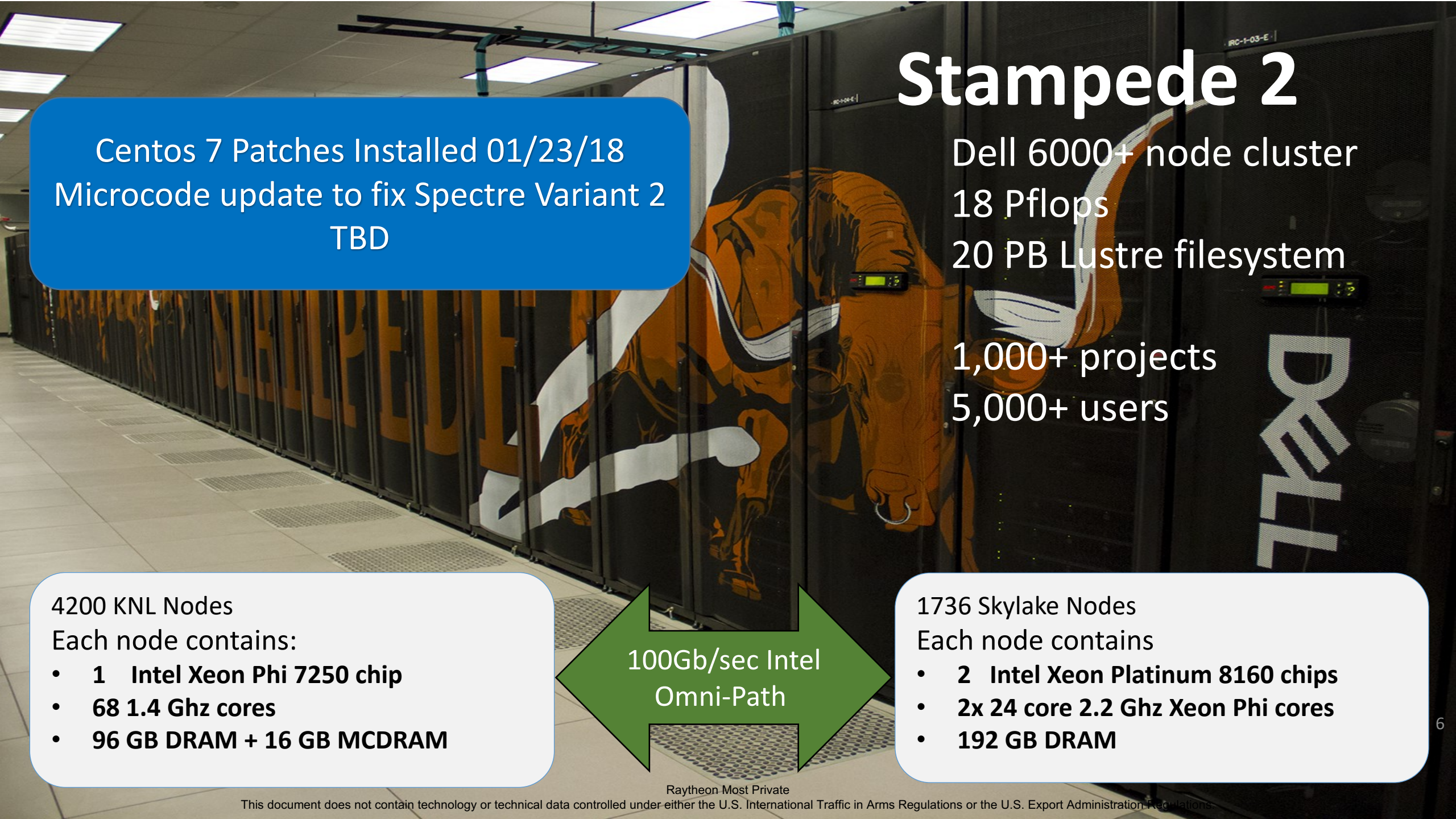
4200 KNL Nodes
Each node contains:

- **1 Intel Xeon Phi 7250 chip**
- **68 1.4 Ghz cores**
- **96 GB DRAM + 16 GB MCDRAM**

100Gb/sec Intel Omni-Path

1736 Skylake Nodes
Each node contains

- **2 Intel Xeon Platinum 8160 chips**
- **2x 24 core 2.2 Ghz Xeon Phi cores**
- **192 GB DRAM**



Stampede 2

Dell 6000+ node cluster
18 Pflops
20 PB Lustre filesystem

1,000+ projects
5,000+ users

Centos 7 Patches Installed 01/23/18
Microcode update to fix Spectre Variant 2
TBD

4200 KNL Nodes
Each node contains:

- **1 Intel Xeon Phi 7250 chip**
- **68 1.4 Ghz cores**
- **96 GB DRAM + 16 GB MCDRAM**

100Gb/sec Intel Omni-Path

1736 Skylake Nodes
Each node contains

- **2 Intel Xeon Platinum 8160 chips**
- **2x 24 core 2.2 Ghz Xeon Phi cores**
- **192 GB DRAM**

Top Applications on Stampede1

Top 6 applications on Stampede 1 from 12/2016 – 11/2017

- VASP -- Ab Initio quantum mechanics
- NAMD -- Molecular dynamics
- Gromacs -- Molecular dynamics
- LAMMPS -- Molecular dynamics
- Chroma -- Quantum chromodynamics
- WRF -- Mesoscale weather forecast

HPC Applications Tested

- WRF -- Mesoscale weather forecast
- GROMACS -- Molecular dynamics
- NAMD -- Molecular dynamics
- GSI -- Meteorological data assimilation

Added due to I/O component

Benchmark data for both KNL and SKX available from pre-kernel patch

At least 3 runs of each were performed and the minimum time or maximum performance was chosen for comparison

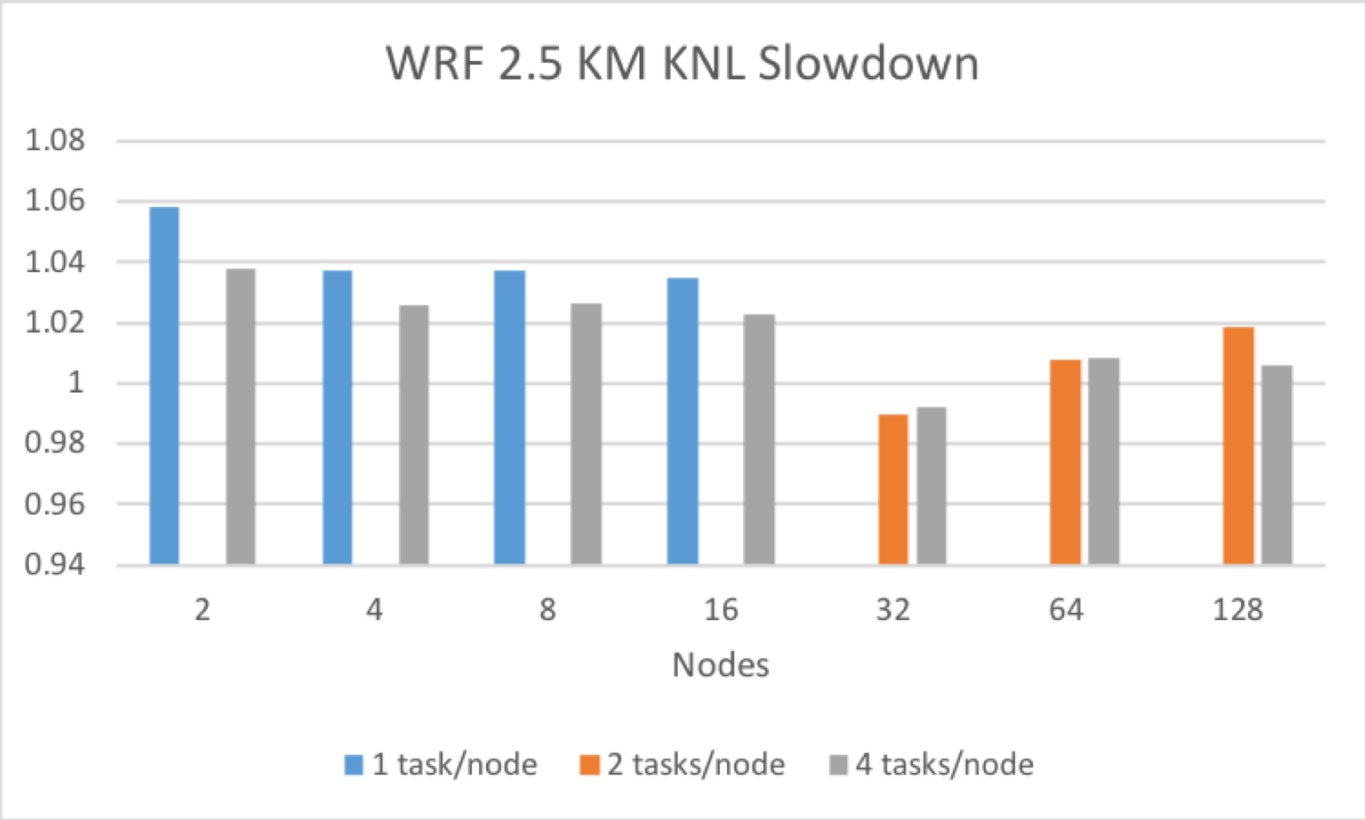
- Except for GROMACS

WRF – Weather Research and Forecasting Model

Version 3.6.1

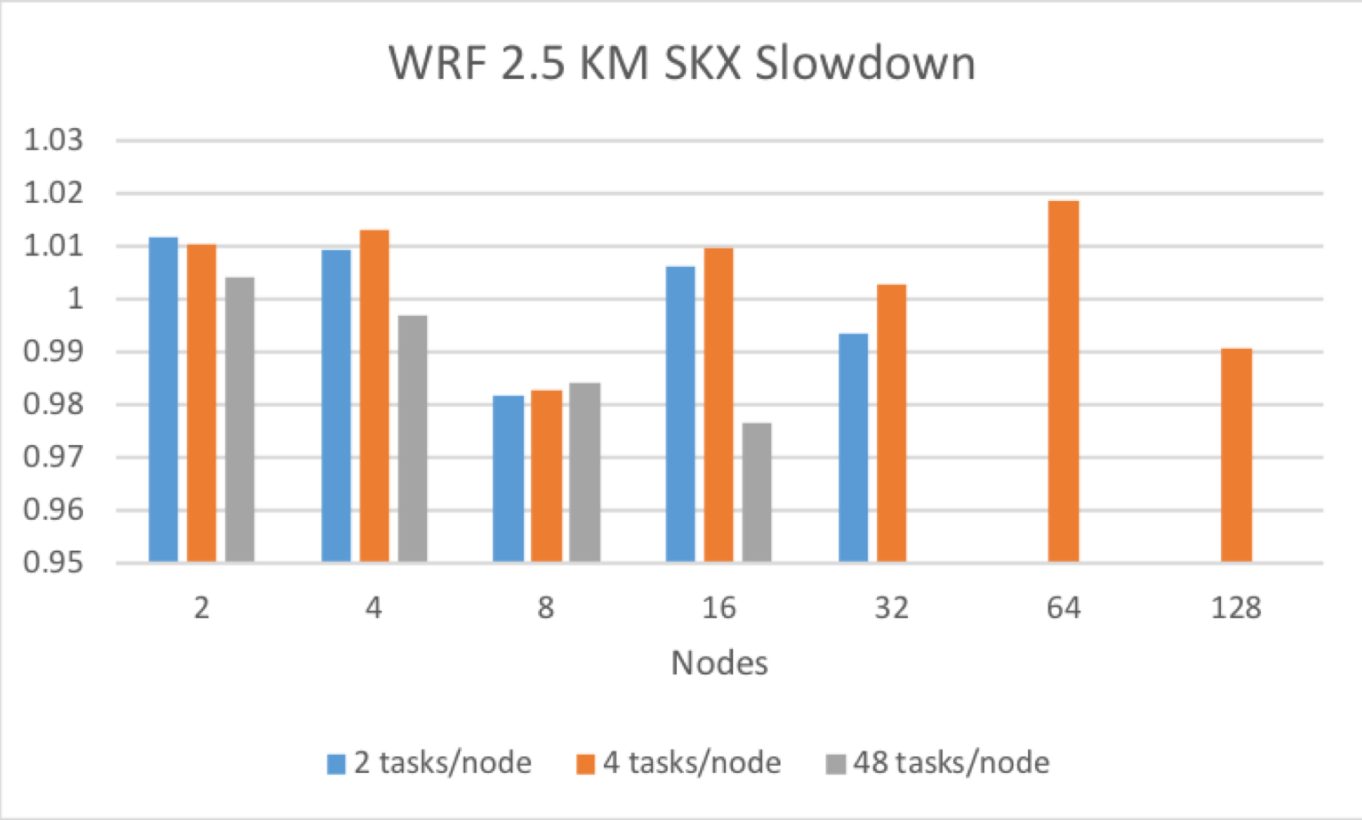
- Finite difference weather code
- 2.5 KM benchmark
- 1 KM test case
- Timings for domain 1 execution – no I/O
- Comparing best time out of 3 runs
  - Statistics limited by what was run before the patch

WRF Results 2.5KM KNL



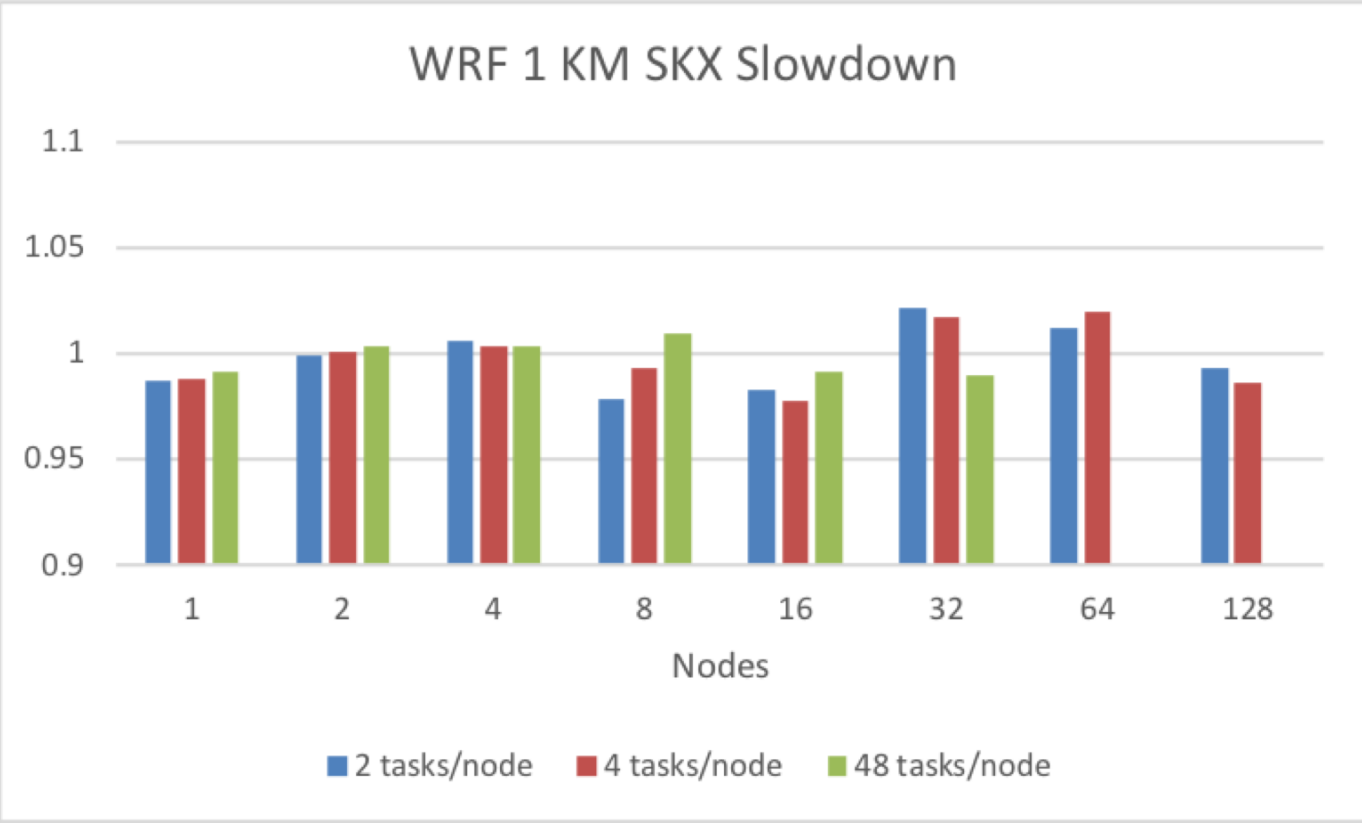
- Slowdown = $\frac{\text{Time after patch}}{\text{Time before patch}}$
- Does not include I/O
- Impact noticeable at 2 nodes -- only 6%

WRF Results 2.5KM SKX



- Does not include I/O
- No noticeable impact
- Data limited to pre-patch runs

WRF Results 1KM SKX



- Domain 1 execution times
- Does not include I/O
- No noticeable impact
- < +/- 3%

Gromacs – Scalar Molecular Dynamics

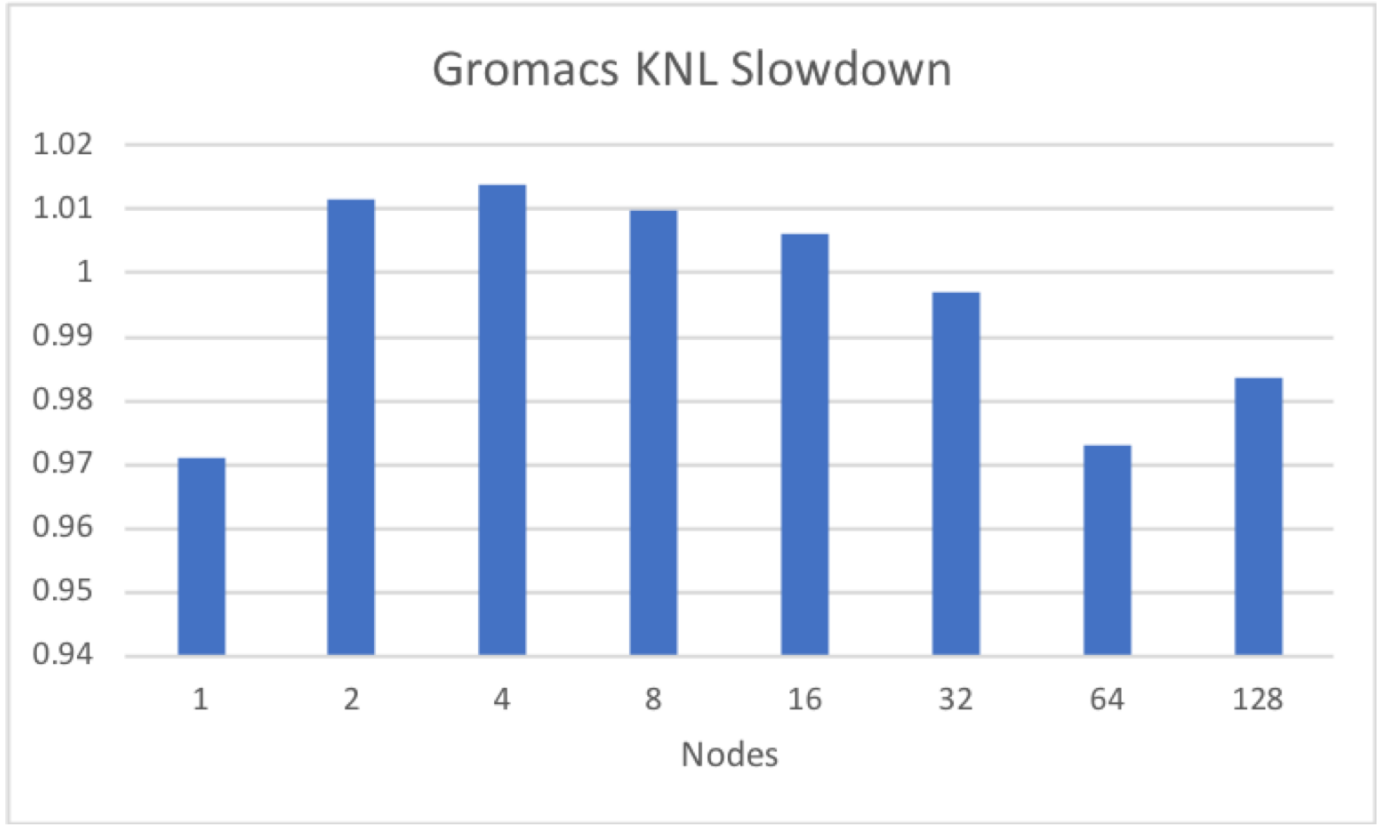
Gromacs Version 2016.3

Built using default stack

- Intel 17.0.1
- Intel MPI 2017.4
- Optimized for KNL & SKX

Pure water test case – 3 million atoms

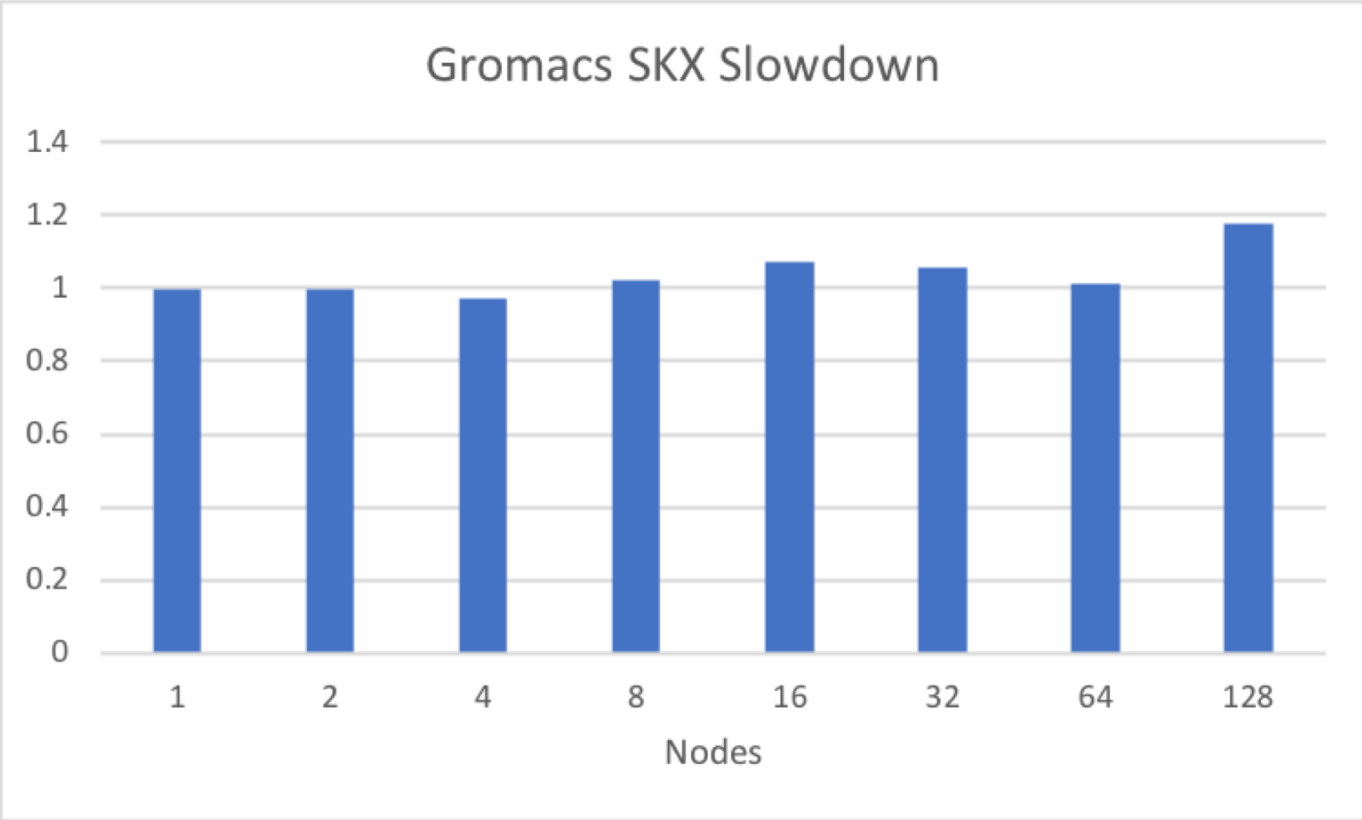
Gromacs Results KNL



Impact barely noticeable
– less than 5%

However
– Only one run from prepatch period

Gromacs Results SKX



No noticeable impact except for 128 nodes

—Parallel efficiency drops to 40%

But, with only 1 prepatch run, statistics are poor

GSI – Gridpoint Statistical Interpolation system

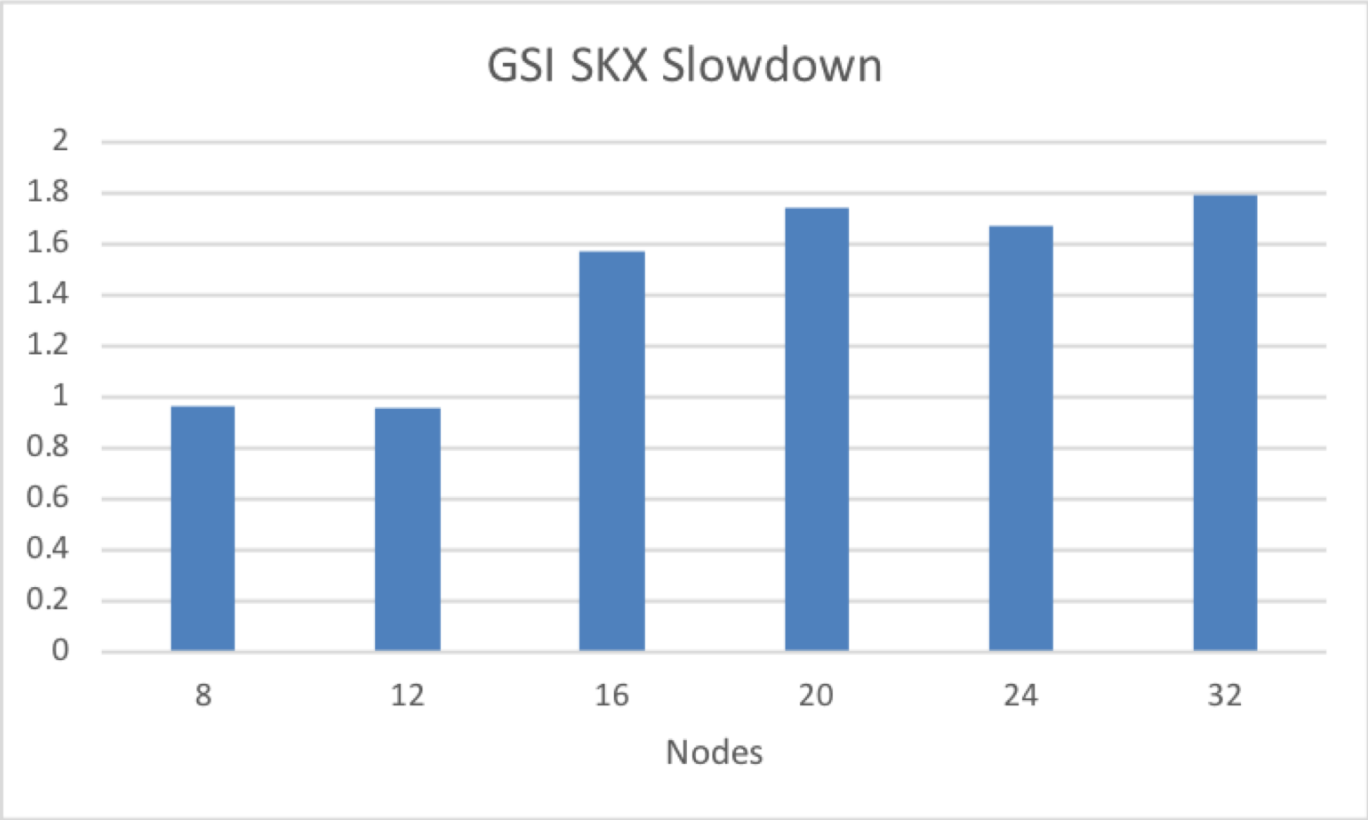
Data Assimilation system used in NOAA Operations

- ~30% walltime spend performing I/O
- Only I/O heavy benchmark with pre-patch results

Comparing best performance of 3 runs

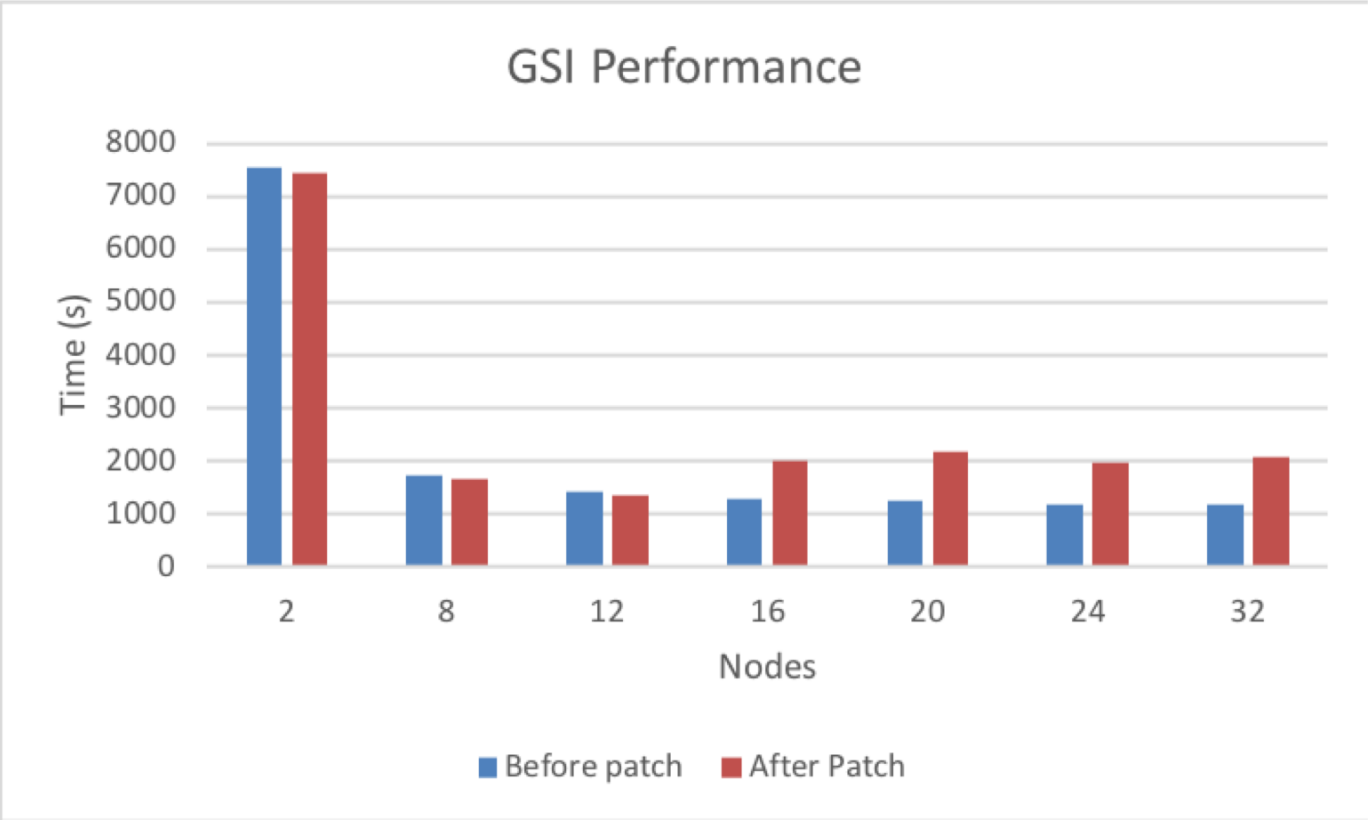
- Statistics limited by what was run before the patch

GSI Results SKX



- Strong effect at higher node counts
- ~30% execution time in I/O
- BUT ….

GSI Performance Results SKX



- Slowdown blows up when GSI quits scaling

NAMD – Scalar Molecular Dynamics

NAMD Version 2.12

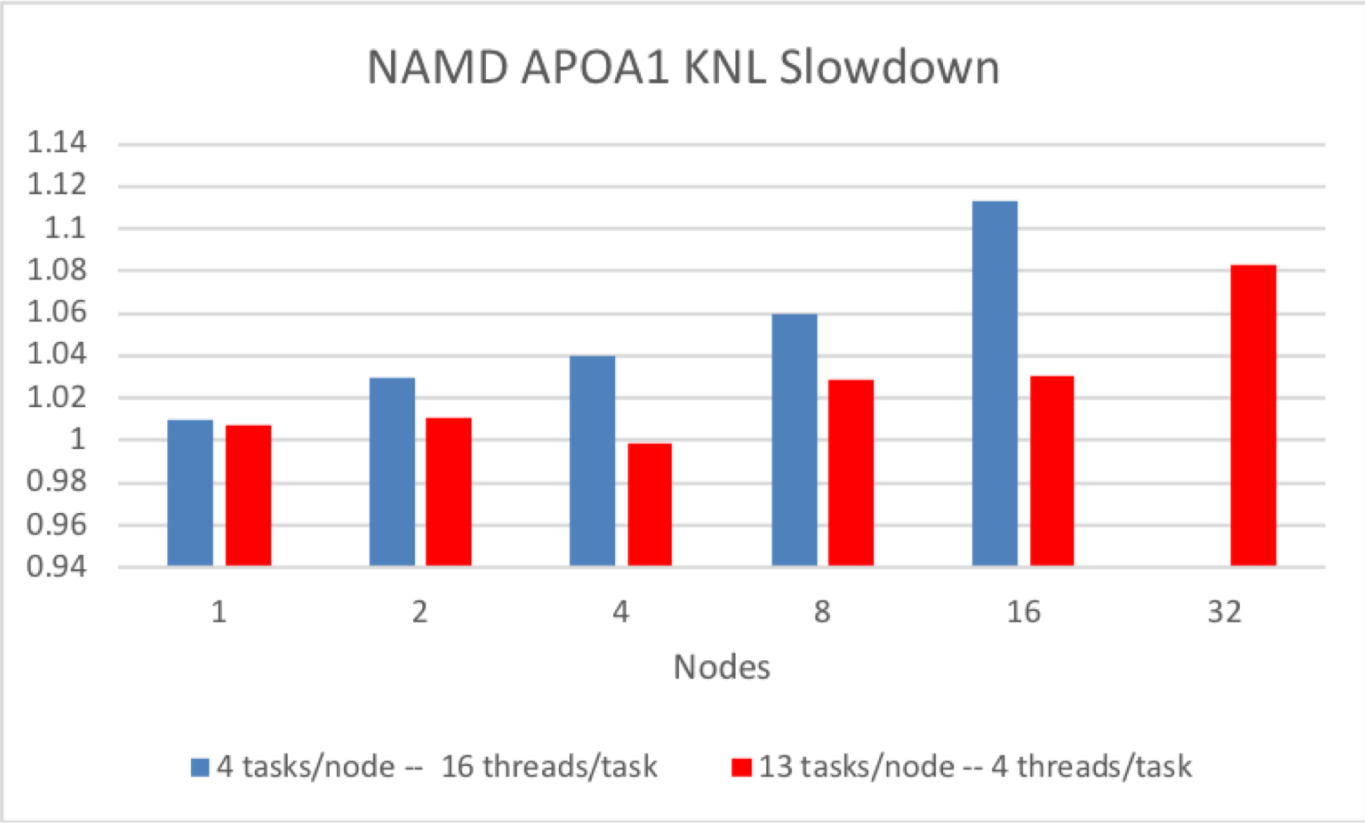
Charm++ Version 6.7.1

* Built using Intel 16.0.3 rather then Intel 17.0.1 due to performance issues with Intel 16

2 test cases

- Task & thread counts chosen for optimal performance
  - STMV-20 -- 1 million atoms
  - APOA1 -- 92k atoms
- Comparing best performance of 3 runs
  - Statistics limited by what was run before the patch

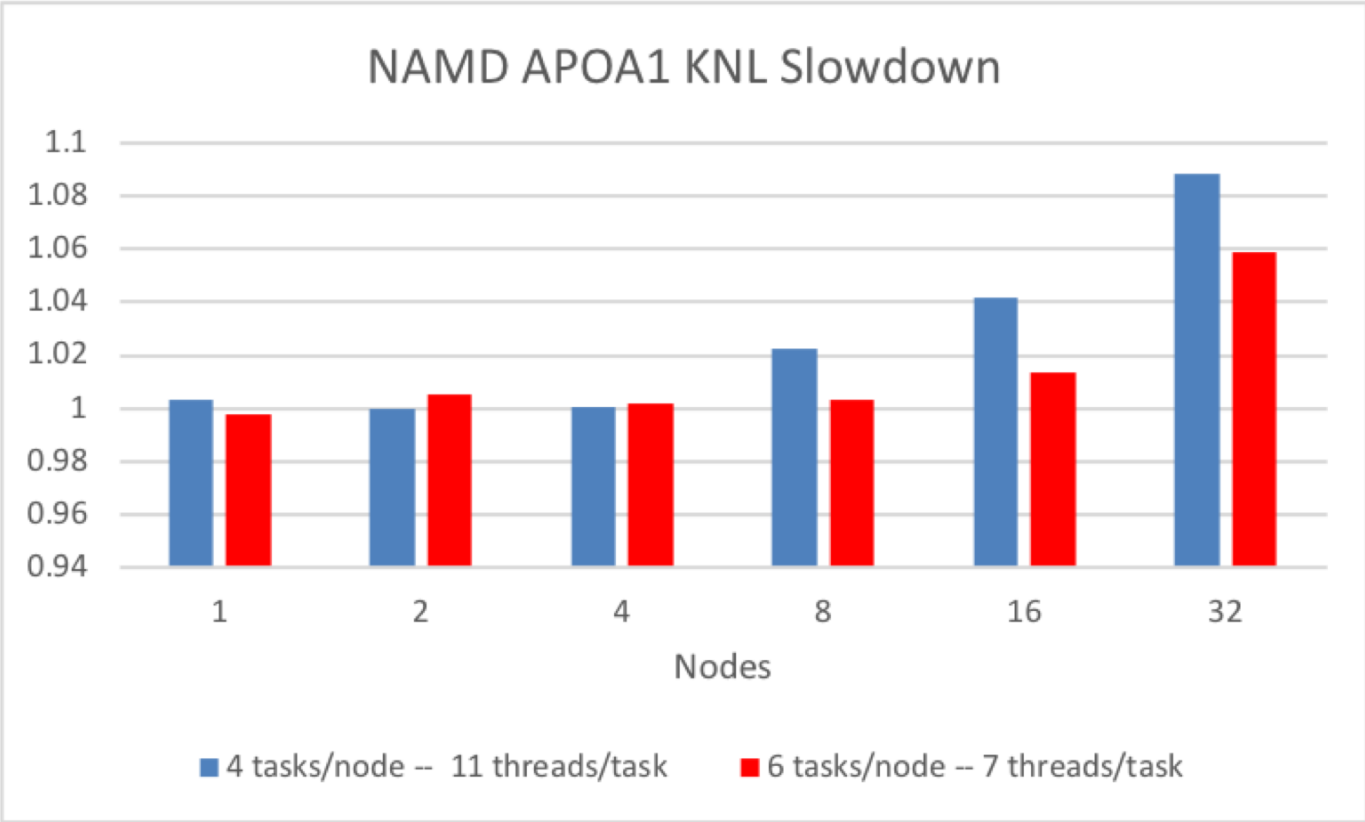
NAMD Results APOA1 KNL



$$\text{Slowdown} = \frac{\frac{\text{ns}}{\text{day}}\text{before patch}}{\frac{\text{ns}}{\text{day}}\text{after patch}}$$

- >1 is slower
- Gets worse as we scale out
- Performance scaling ends at 16 nodes
- Counterintuitive that fewer tasks exhibits

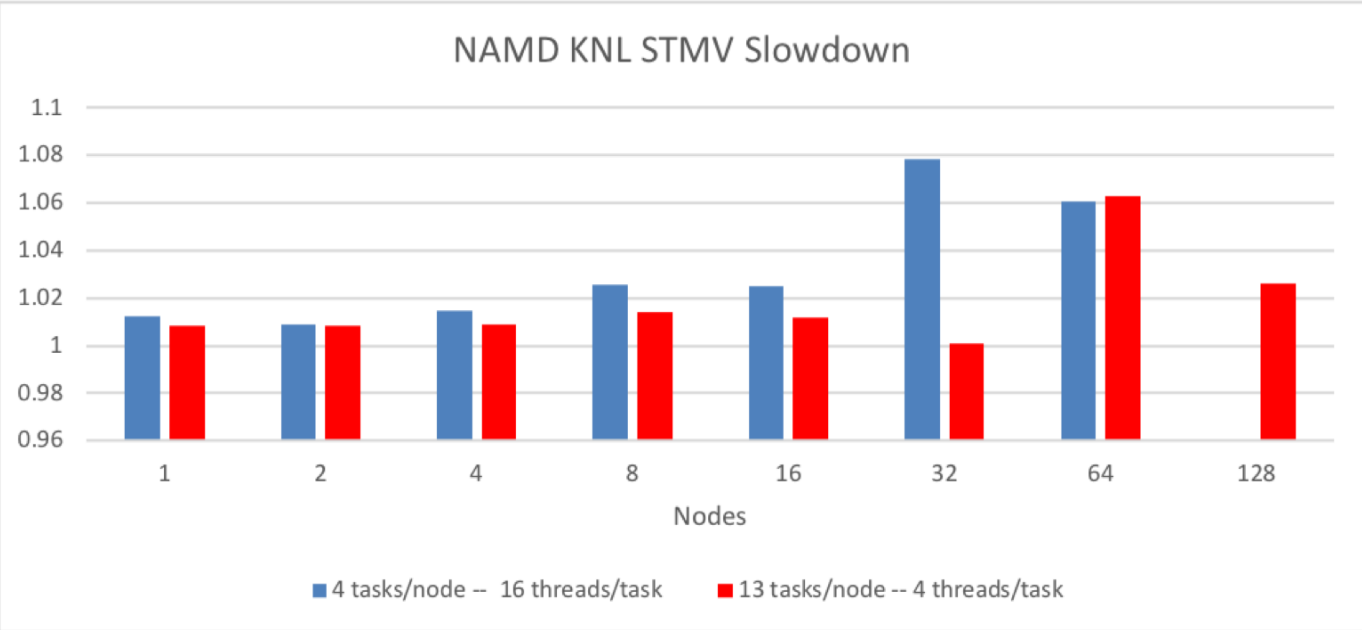
NAMD Results APOA1 SKX



Performance still drops off as we scale out

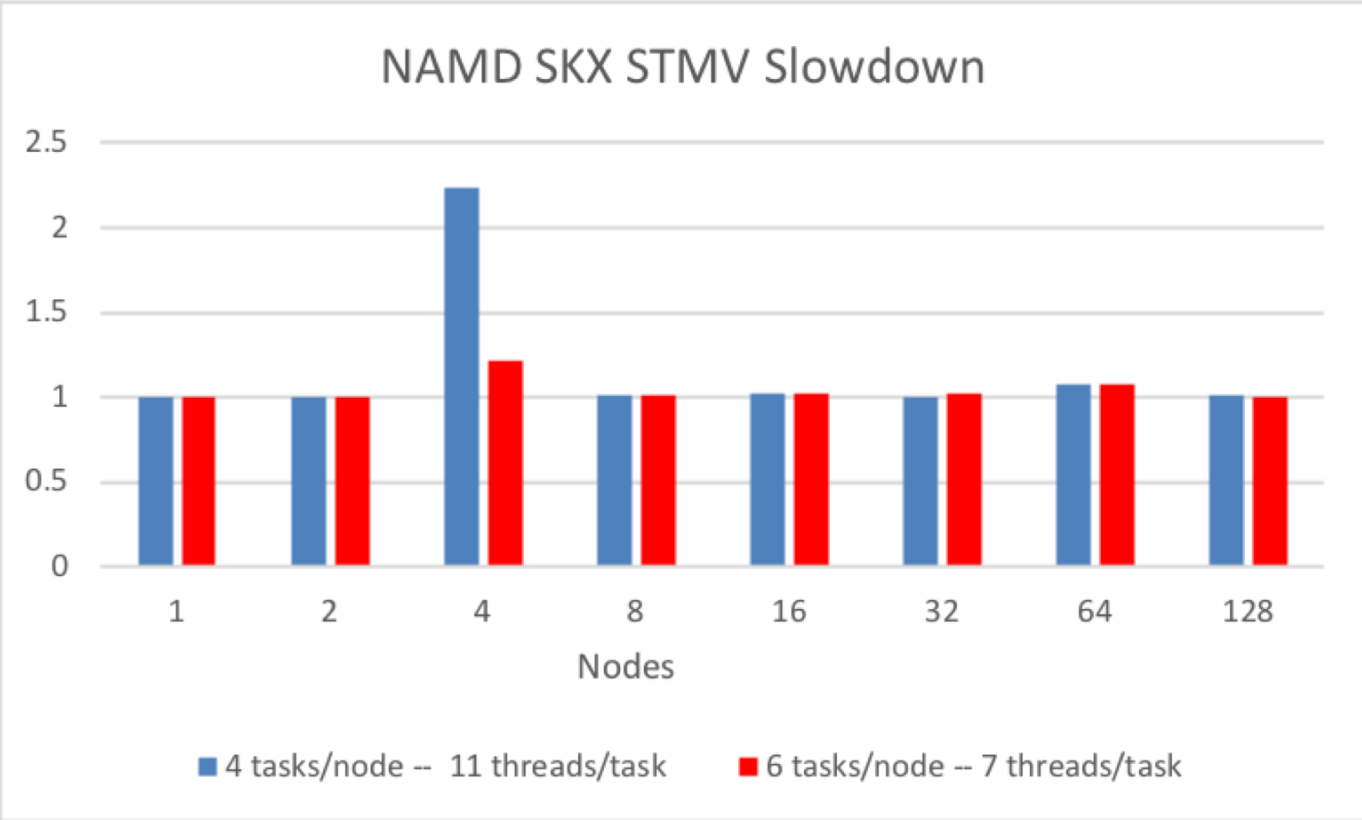
- Not as bad as KNL
- Scaling stops at 16 nodes

NAMD Results STMV KNL



- Slowdown = $\frac{\frac{\text{ns}}{\text{day}}\text{before patch}}{\frac{\text{ns}}{\text{day}}\text{after patch}}$
- >1 is slower
- Slowdown gets worse as we scale out
- Performance scaling ends
  - 16 nodes – 4 tasks/node
  - 32 nodes – 13 tasks/node

NAMD Results STMV SKX

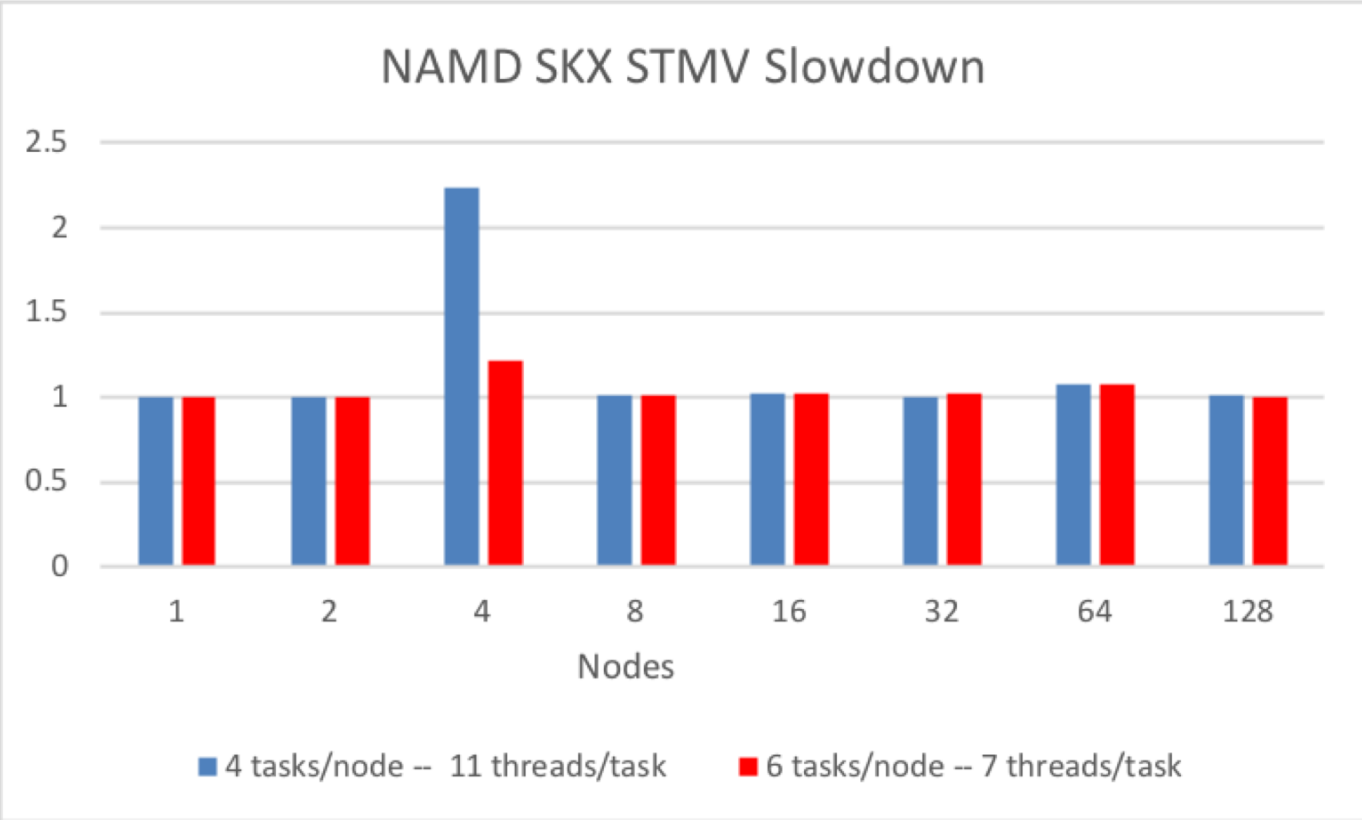


One really large outlier
One smaller outlier

- Both are at 4 tasks/node

What's going on?

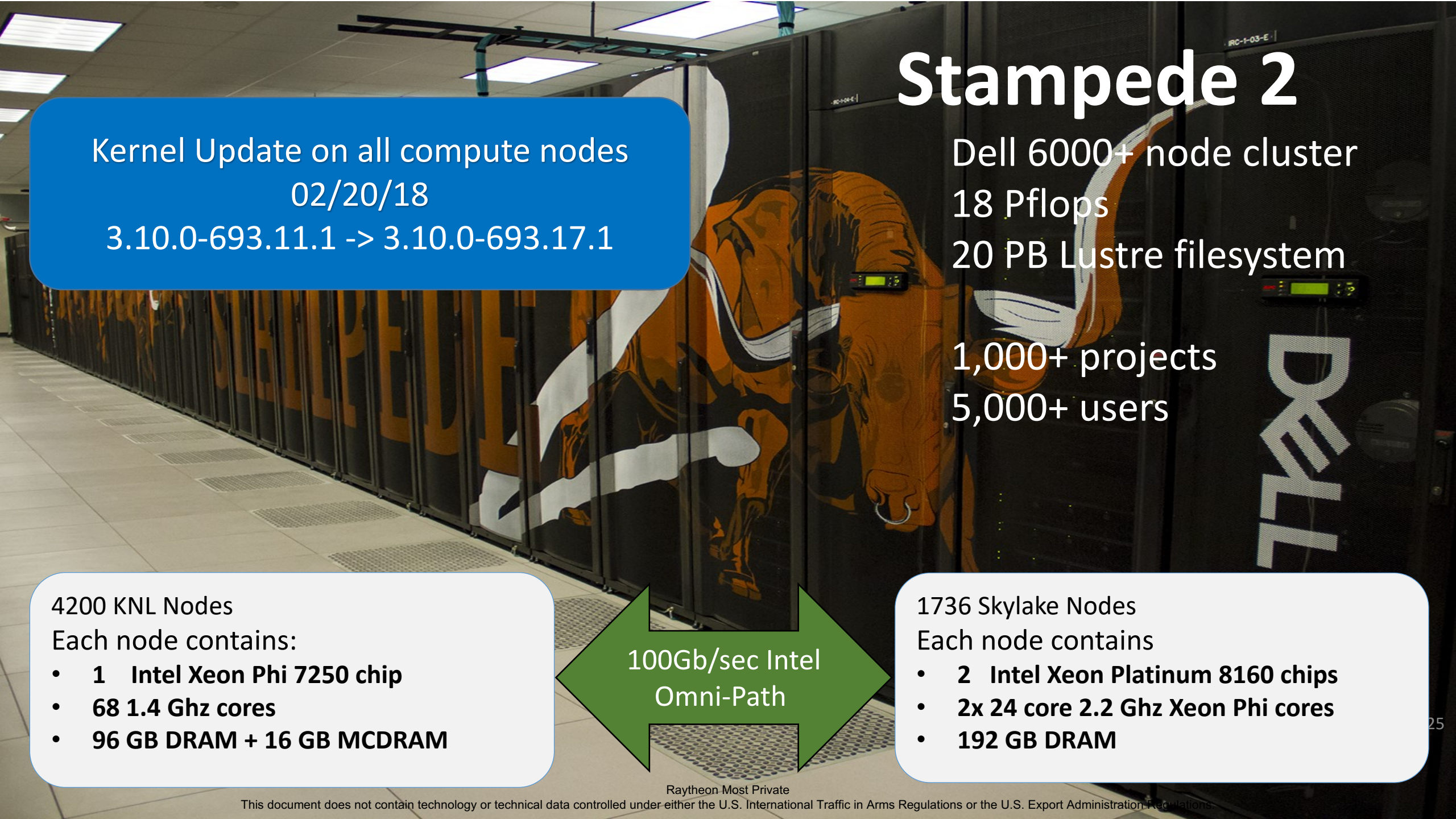
NAMD Results STMV SKX



Repeated results with 10 runs on different node sets

Profiling attributed all the extra time to system time

Planned to test this on node subset with kernel patches disabled



Stampede 2

Dell 6000+ node cluster
18 Pflops
20 PB Lustre filesystem

1,000+ projects
5,000+ users

Kernel Update on all compute nodes
02/20/18
3.10.0-693.11.1 -> 3.10.0-693.17.1

4200 KNL Nodes
Each node contains:

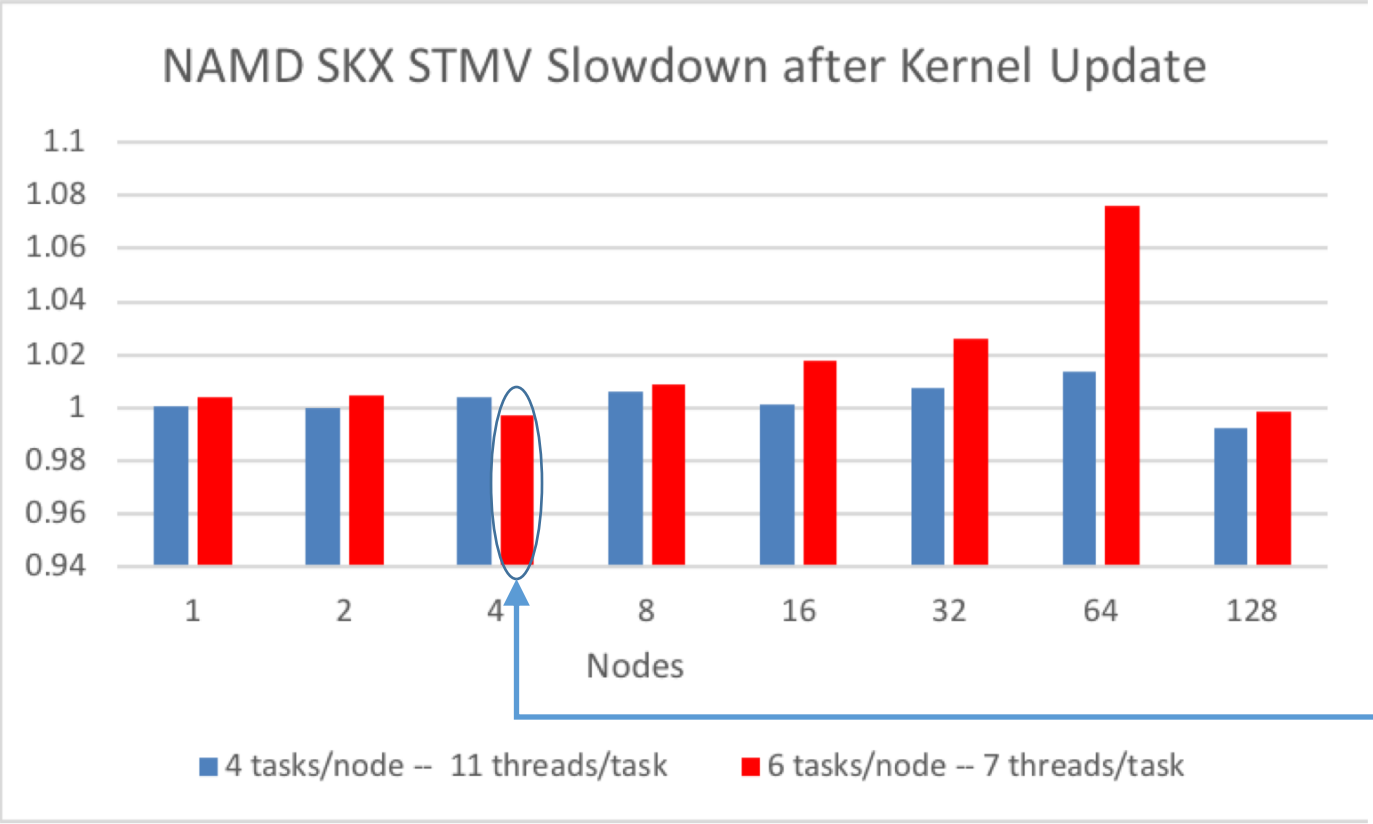
- **1 Intel Xeon Phi 7250 chip**
- **68 1.4 Ghz cores**
- **96 GB DRAM + 16 GB MCDRAM**

100Gb/sec Intel Omni-Path

1736 Skylake Nodes
Each node contains

- **2 Intel Xeon Platinum 8160 chips**
- **2x 24 core 2.2 Ghz Xeon Phi cores**
- **192 GB DRAM**

NAMD Results STMV SKX after Kernel Update



Reran 4 tasks/node on all node counts

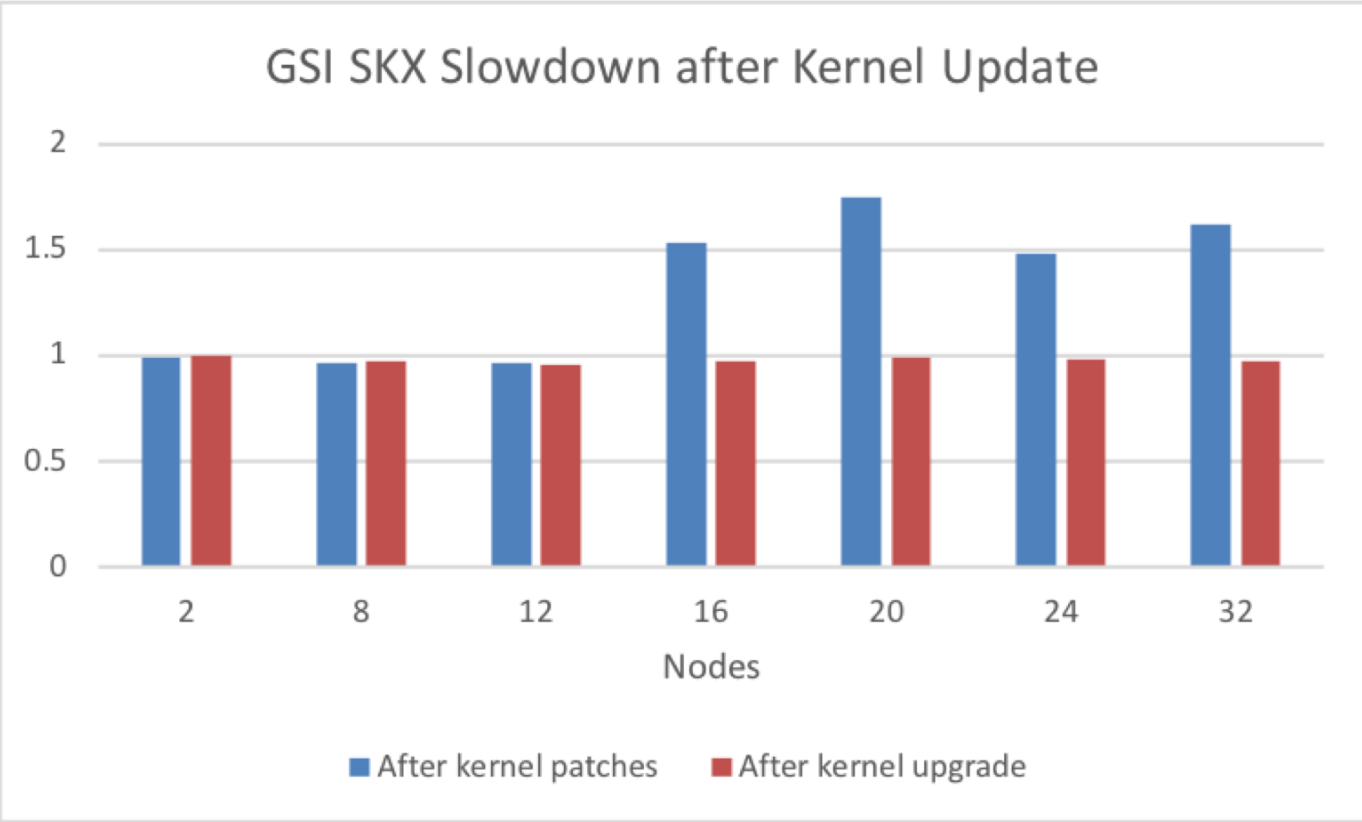
- Performance back to normal

Reran 6 tasks/node on 4 nodes only

- 20% slowdown disappears!

???

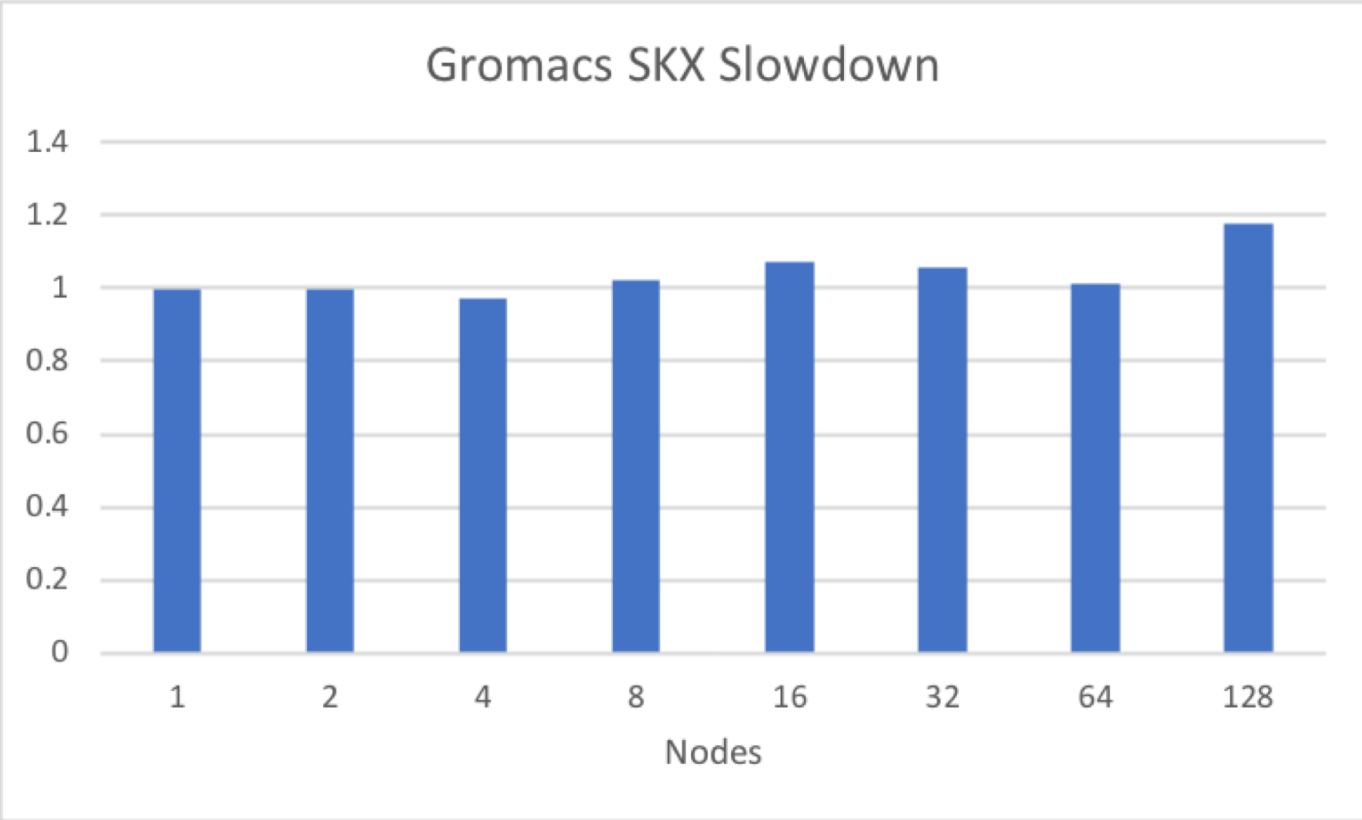
GSI Results SKX after Kernel Update



Reran on SKX after update

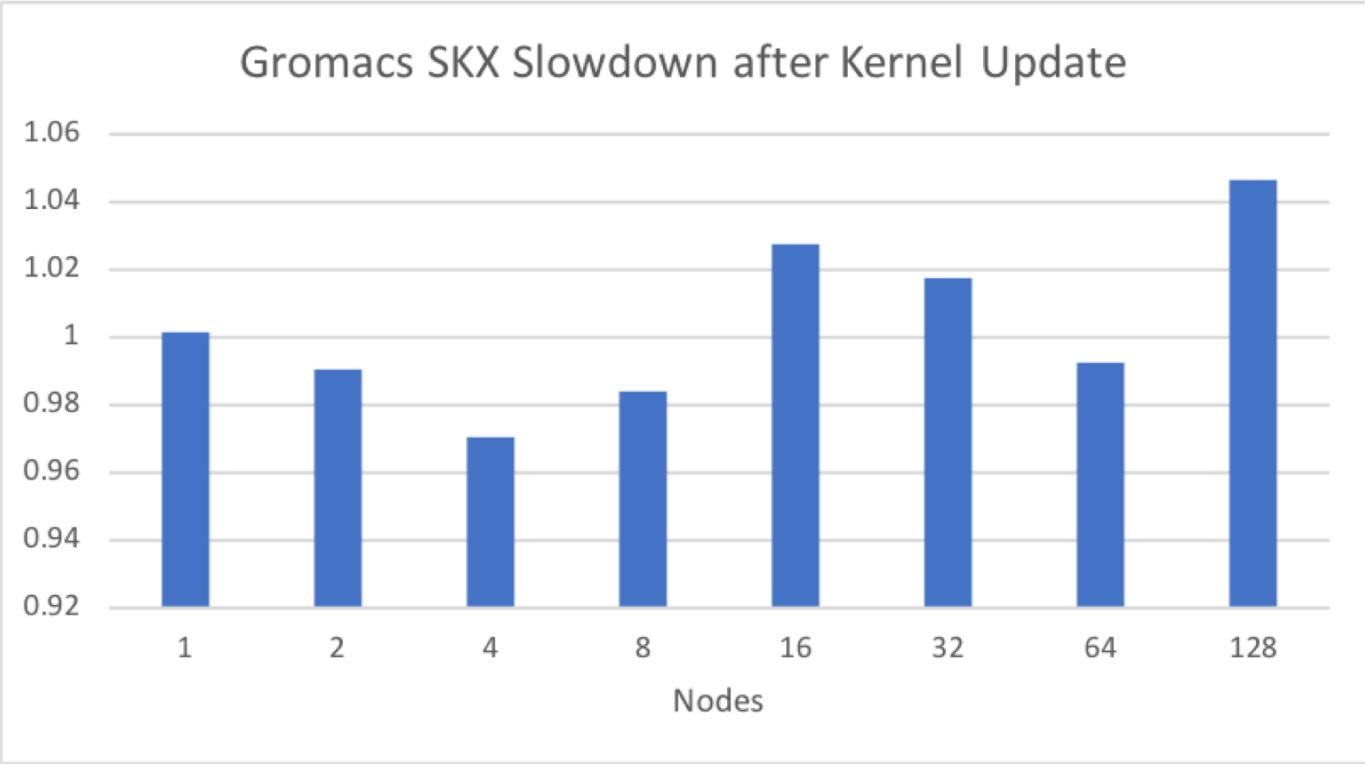
- Performance back to normal

Gromacs Results SKX



Original gromacs results had an outlier with a 20% slowdown

Gromacs Results SKX after Kernel Update



Outlier disappears at 128 nodes

Like magic!

Overall Analysis

- Possible issues with MPI and I/O overhead for small ops
- Doesn’t cause much impact for well designed codes
- KNL/SKX impact < 5% in most cases

Something changed from kernel

3.10.0-693.11.1 -> 3.10.0-693.17.1

Overall Analysis

- Possible issues with MPI and I/O overhead for small ops
- Doesn’t cause much impact for well designed codes
- KNL/SKX impact < 5% in most cases

Something changed from kernel

3.10.0-693.11.1 -> 3.10.0-693.17.1

Nothing to see here

Move along

But . . .

Spectre microcode fix is coming!

Thanks!