Evaluation of SMP Shared Memory Machines for Use With In-Memory and OpenMP Big Data Applications

Andrew J. Younge [*], Christopher Reidy [†], Robert Henschel [*], Geoffrey C. Fox [*]

[*]

School of Informatics and Computing
Indiana University
901 E. 10th St, Bloomington, IN USA
{ajyounge, henschel, gcf}@indiana.edu

[†] Research Computing
University of Arizona
1077 N. Highland Ave, Tucson, AZ USA
chrisreidy@email.arizona.edu

Agenda

- Introduction
- Motivation
  - Thread parallelism vs large memory
- Two SMP Machines:
  - ScaleMP vSMP @ Indiana University
  - SGI UV 1000 @ University of Arizona
- Performance Evaluation
  - SPEC OpenMP benchmarks
  - Trinitiy De-novo Assembly
- Discussion

Introduction

- Natural desire for larger, more powerful computing components
- Need for systems larger than a single node/board/chip can provide
- Distributed memory systems fill this role, but can present a larger barrier of entry
  - Require special programming models
  - Code redesign
  - Explicit parallelism
- Want the performance of distributed memory with the ease of a single resource

Symmetric Multiprocessing

- Desire to treat distributed memory systems like a single, unified computing resource
- Centralized shared memory under a single unified OS with many processing units
  - Historically SMP was multi-socket, now multi-node
  - Operate across a system bus
  - NUMA
- Current implementations have complex cache coherency mechanisms

SMP Use Cases

Thread Parallelism

- Applications can leverage more threads than what is commonly available from a single computing resource
- Easy scaling to SMPs given today's multi-core systems
- Programming models include OpenMP, Pthreads, Cilk, etc

Large Memory

- Amount of main memory needed can be more than what a single system can support
- Allows for fast DRAM to be used instead of disk
- Scaling databases, big data applications, in-memory data management

The goal is to avoid re-writing applications from scratch but still scale beyond single node performance

Role of SMPs for Big Data

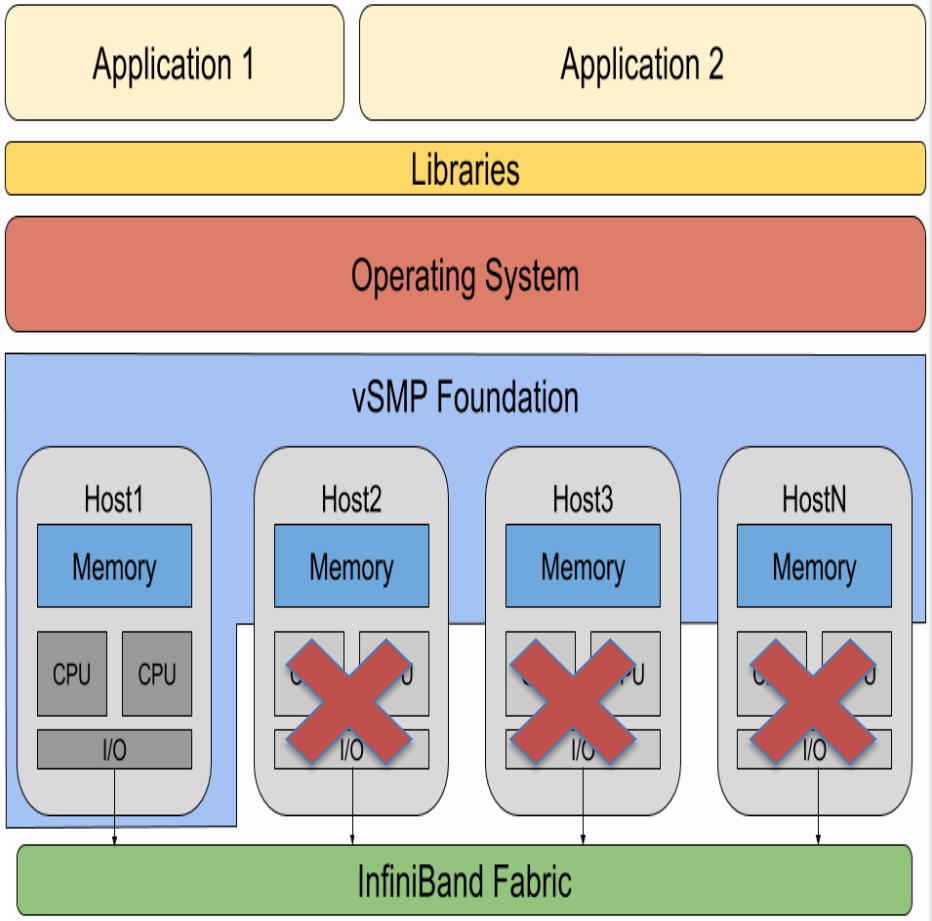
- There are many efforts that are not able to take advantage of distributed memory systems
  - Legacy code too hard to update
  - Proprietary code unable to be modified
  - Development effort exceeds application usability
- SMPs enable the utilization of distributed systems in an environment that appears as a single entity
  - 100s of cores, TBs of RAM
- However SMP performance is often nondeterministic and lacking clarity for use with big data applications

ScaleMP vSMP

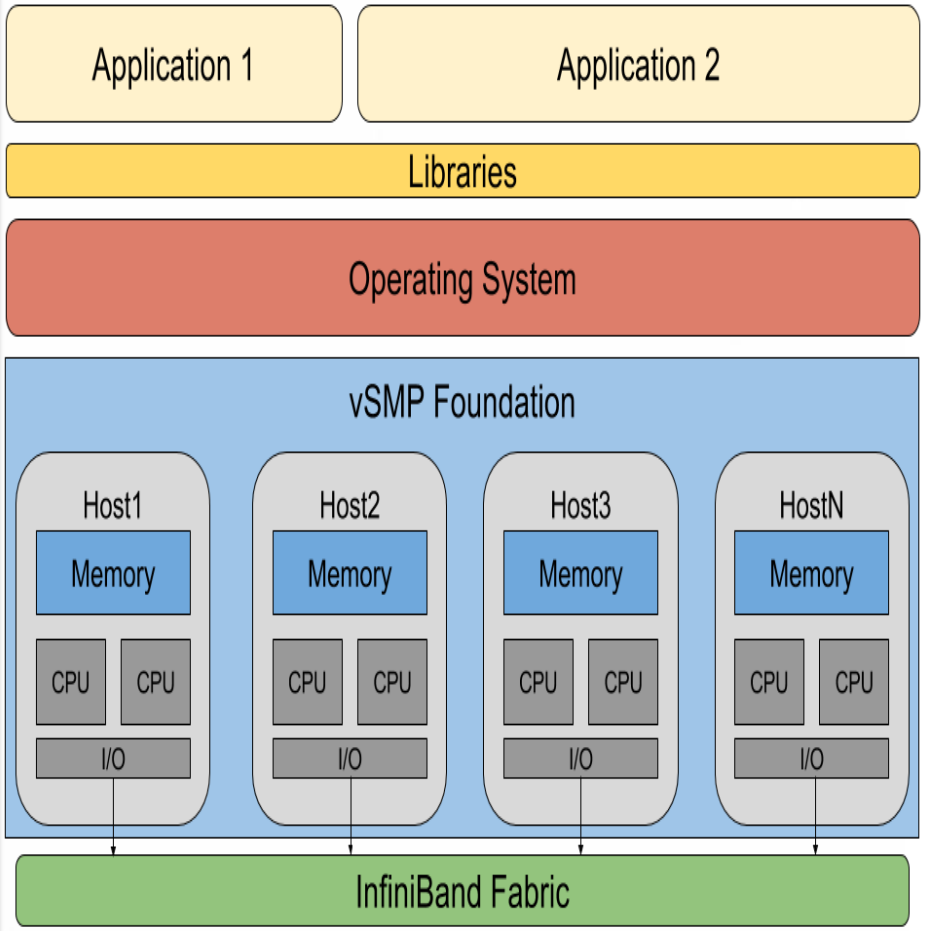
- Aggregates commodity x86 servers to create a single, virtualized system.
  - Looks like a large VM with 1 OS with all cores & memory available
  - Can scale to 128 nodes (1024 cores, 128TB RAM)
- vSMP Foundation
  - Runs via master/slave nodes
  - Leverage RDMA via InfiniBand interconnect
- Data locality, CPU organization handled by vSMP and OS kernel scheduling & tools

vSMP Expansion Modes

Memory Expansion



System Expansion



Echo @ IU

- Echo is the vSMP installation as part of the NSF FutureGrid project (now FutureSystems) at Indiana University
- 16 nodes, 192 cores, 5.6TB RAM
  - Mellanox CX3 QDR InfiniBand
- Purchased in late 2013, based on Intel Xeon Sandy-Bridge

ScaleMP™

SGI Ultraviolet (UV)

- SGI has been providing SMP machines for many years
  - Design dates back to original Stanford DASH computer
- Creates global shared memory system across many blades
- Developed specialized NUMAlink interconnect
  - Global Register unit for Cache Coherency
  - Active Memory Unit for atomic operations
  - Socket based Intel QPI interface

SGI UV

- SGI UV installation at University of Arizona
- UV 1000 – 58 nodes at 1776 cores, ~10 Tflops
  - 16 nodes reserved exclusively for our experimentation
- Purchased in 2011, based on Intel Xeon Westmere cores

Hardware Comparison

	ScaleMP	SGI UV1000
Machine Name	Echo	UV
Nodes	16	16
NUMA Nodes	32	32
Cores	192	256
System RAM	5.6 TB (6.0 TB)	1.28 TB
CPU Type	Intel x86_64	Intel x86_64
CPU Model	Xeon E5-2640	Xeon E7-8837
CPU Family	SandyBridge	Westmere
CPU Base Frequency	2.5 Ghz	2.66 Ghz
CPU Max Frequency	3.0 Ghz	2.8 Ghz
CPU Cores	6	8
Cache Size	15 MB	24 MB
Interconnect	Mellanox QDR InfiniBand	SGI NUMAlink 5
IC Bandwidth	40 Gbps	15 Gbps
OS Type	RedHat EL6 Linux	RedHat EL6 Linux
OS Kernel	2.6.32-220.23.1.vSMP.el6.x86_64	2.6.32-358.6.2.hz100.el6.x86_64
SMP Version	vSMP 6.0.135.46	SGI Accelerate 1.6 SGI Foundation 2.8
Compiler	Intel 2013.sp1.2.144	Intel 2013.5.192

TABLE I
HARDWARE CONFIGURATION

Other SMP Solutions

- We are evaluating 2 SMP solutions, but others exist
  1. BullX SuperNode S6000 - 288 cores, 24TB RAM, IB
  2. “Fat node” single server solution - ~2TB
  3. Intel Xeon Phi - many cores, little RAM
  4. IBM POWER8 8-way SMP

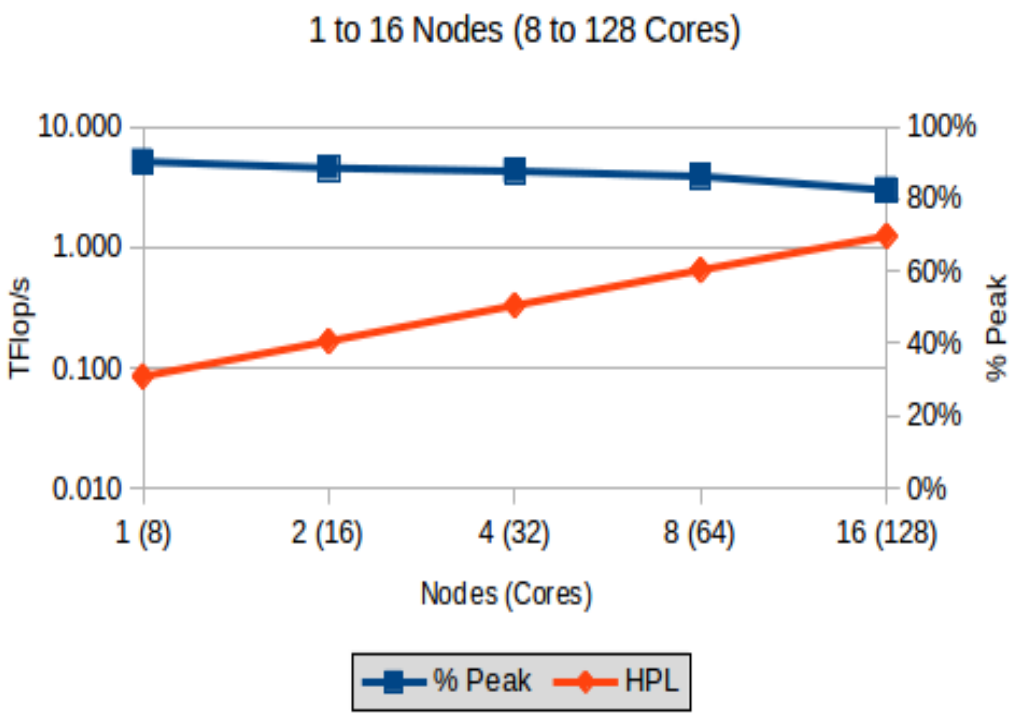
Performance Evaluation

- Evaluating the performance of two dissimilar SMP machines: vSMP and SGI UV
- Apples-to-apples comparison not possible, but want make best effort to keep as equal as possible
  - Same node count, similar OS & kernels, etc
- Look at our two use cases: Task-level parallelism and in-memory big data apps

MPI on SMP?

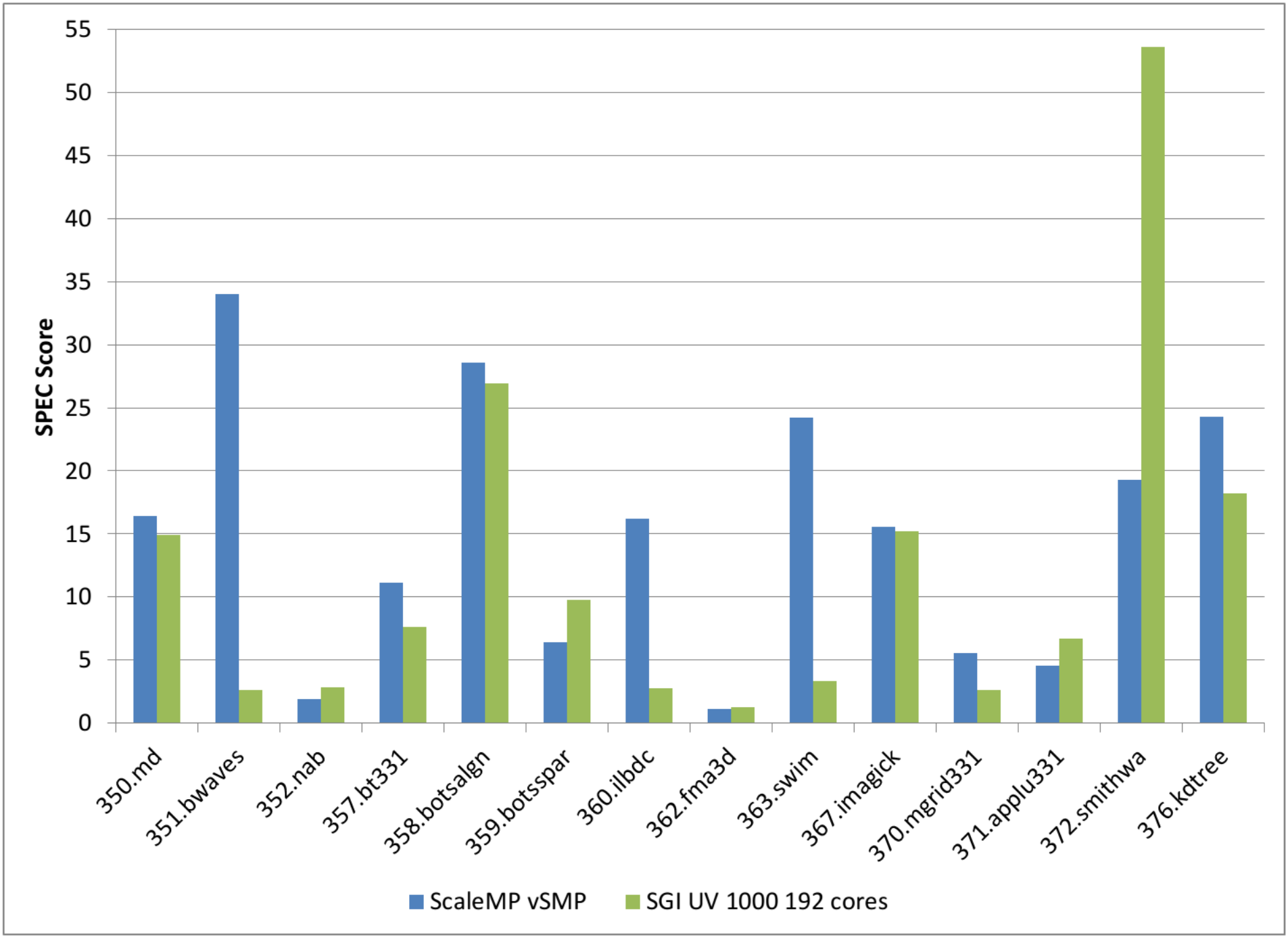
- While MPI not the target for SMP machines, both ScaleMP and SGI support MPI apps
  - ScaleMP provides MPICH2 build
  - SGI has specialized MPI offload engine
- Overall, MPI performance on both SGI UV and vSMP are good
  - vSMP: 82.4% peak

HPL Performance on “India” test ScaleMP vSMP system (16 nodes)



SPEC OMP 2012

- Standard Performance Evaluation Corporation (SPEC) is a consortium for systems benchmarking
- Developed SPEC OMP 2012 as an OpenMP benchmark suite
  - Based on original OMP 2001 suite
  - Leverages HPC simulation applications
  - Ranges from MD, CFD, LU factorization, alignment…
- Uses a reference system for result comparison
  - Sun FireX4140 with two AMD Opteron 2384 CPUs (quad core)


SPEC Score
55
50
45
40
35
30
25
20
15
10
5
0
350.md
351.bwaves
352.nab
357.bt331
358.botsalgn
359.botsspar
360.ilbdc
362.fma3d
363.swim
367.imagick
370.mgrid331
371.applu331
372.smithwa
376.kdtree
ScaleMP vSMP
SGI UV 1000 192 cores

SPEC OMP Results

- Immediately, no simple story across 14 OMP benchmarks
  - 352.nab, 362.fma3d, 370.mgrid, and 371.applu331 do not scale well
  - 350.md, 358.botsalgn, 372.smithwa, and 376.kdtree, scale well on both systems
- SPEC Score: vSMP = 10.4 and SGI = 7.01
  - 9 benchmarks comparable between both machines Adjusted score: 7.65 vs 7.44 (resp)
  - 4 benchmarks where vSMP at least 2x better than SGI
  - 1 benchmark where SGI 2.7x faster than vSMP

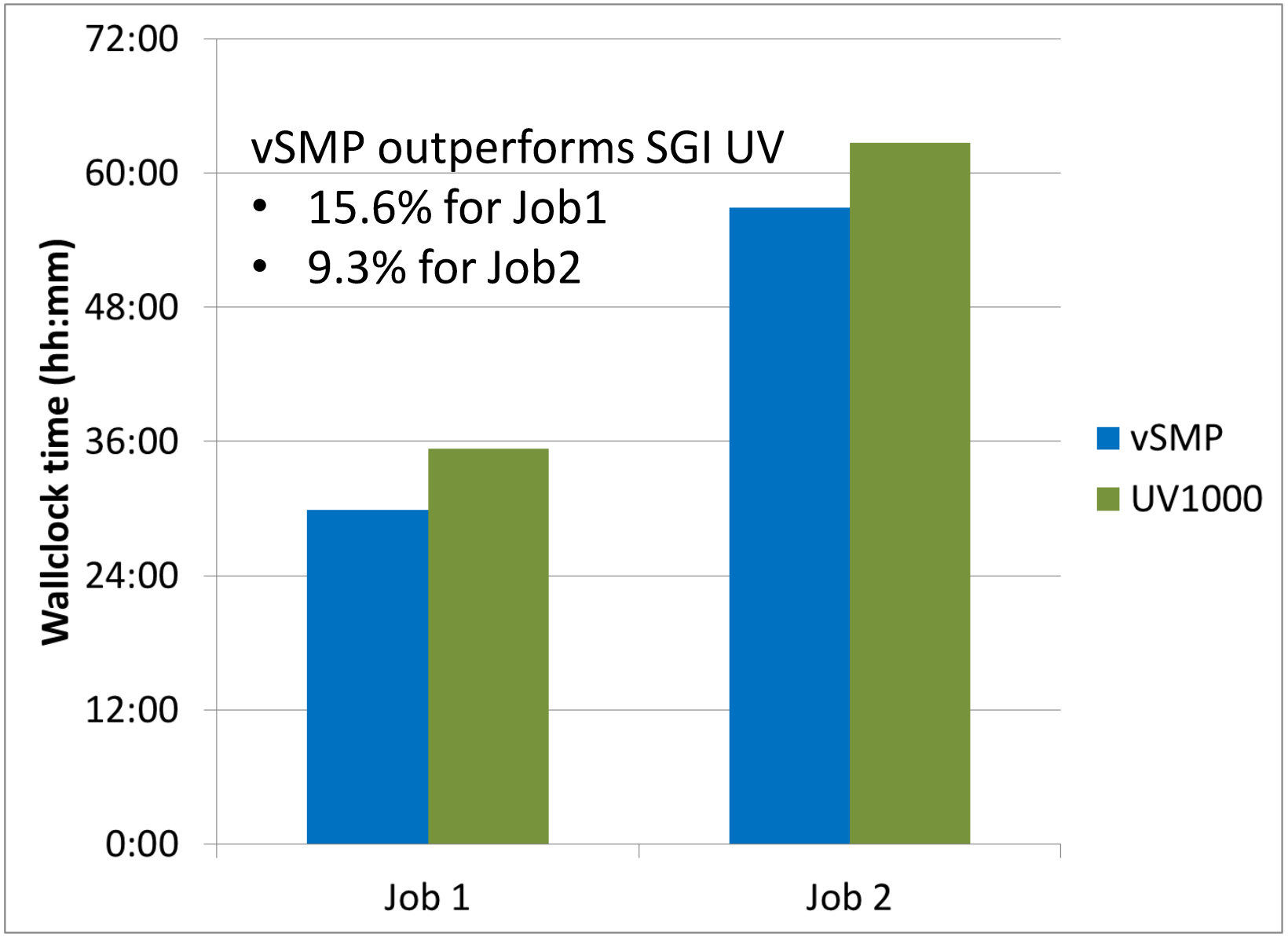
SPEC OMP on Architectures

- Need to consider CPU architecture differences when comparing scores
  - SGI UV is E7-8837 Westmere, 8-cores/socket
  - vSMP is E5-2640 Sandy-Bridge, 6-cores/socket
- Looking at SPEC CPU 2006, see ~17% raw difference between CPUs
  - E7-8837 = 39.3 vs E5-2640 = 46.0
- However, 17% CPU improvement does not explain the 1.5x difference between vSMP and SGI
  - 5.8x difference for large memory footprint OMP benchmarks
  - 40Gbs QDR IB vs 7.5Gbs NUMAlink5
  - Difference is in the Interconnect bandwidth, exacerbated with benchmarks with high memory footprints.

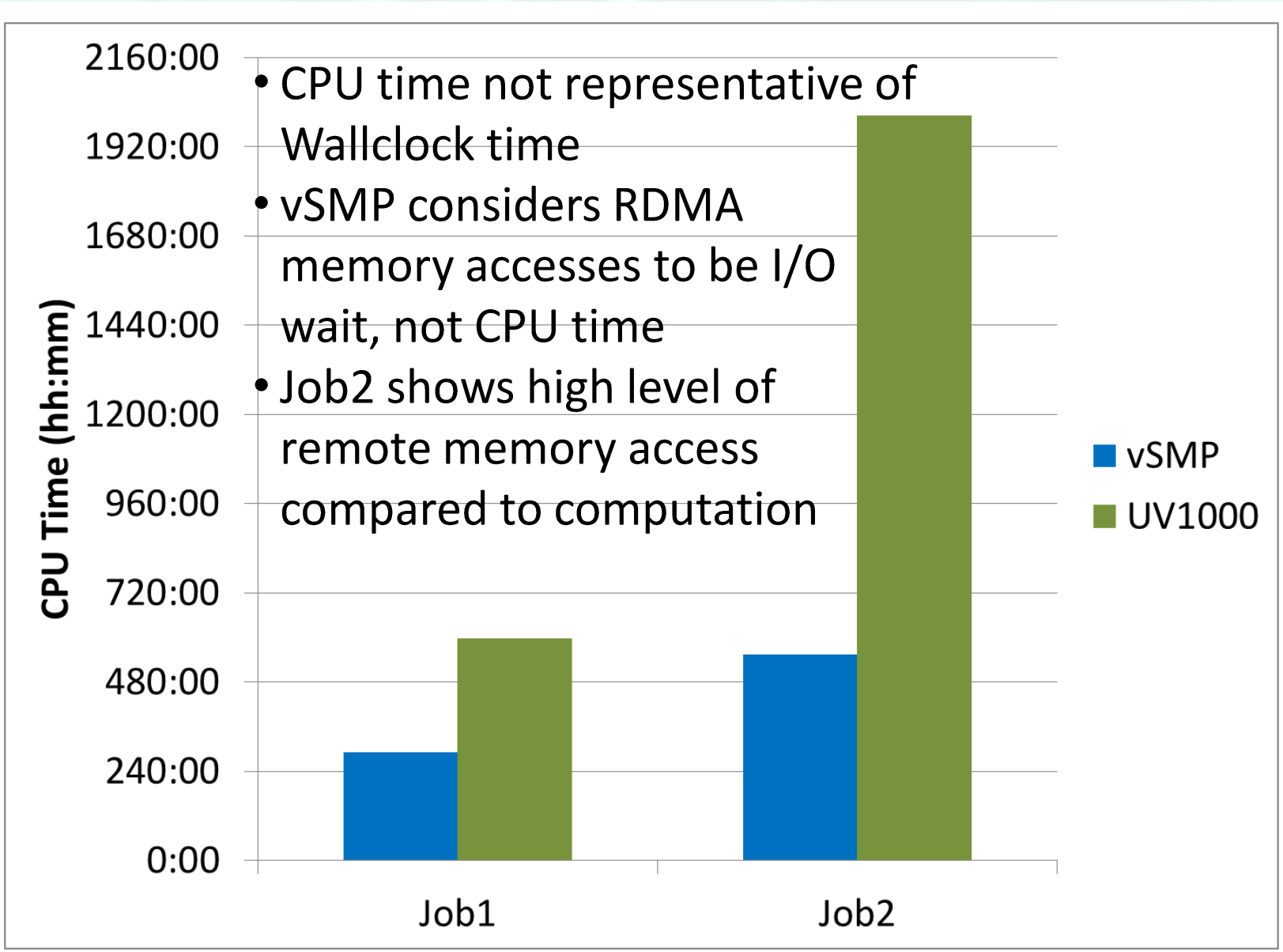
Trinity de novo Assembly

- A transcriptome is the set of all RNA molecules that are transcribed in a cell
  - Help identify genes, alternative splicing in genes, and differential expression of genes in distinct cell populations
- RNA-Seq produces millions of short RNA sequence reads
  - Typically, these huge data sets are down-sampled to produce a smaller data and computed on commodity systems
- Assembled full RNASeq data set of 370 million reads using the Trinity assembly package.
- Two job types based on *Crangon crangon* (Shrimp)
  - gathered from an Illumina RNAseq dataset
  - 100bp reads, ~50x coverage
- Job1: only R1 reads
- Job2: R1 and R2 reads

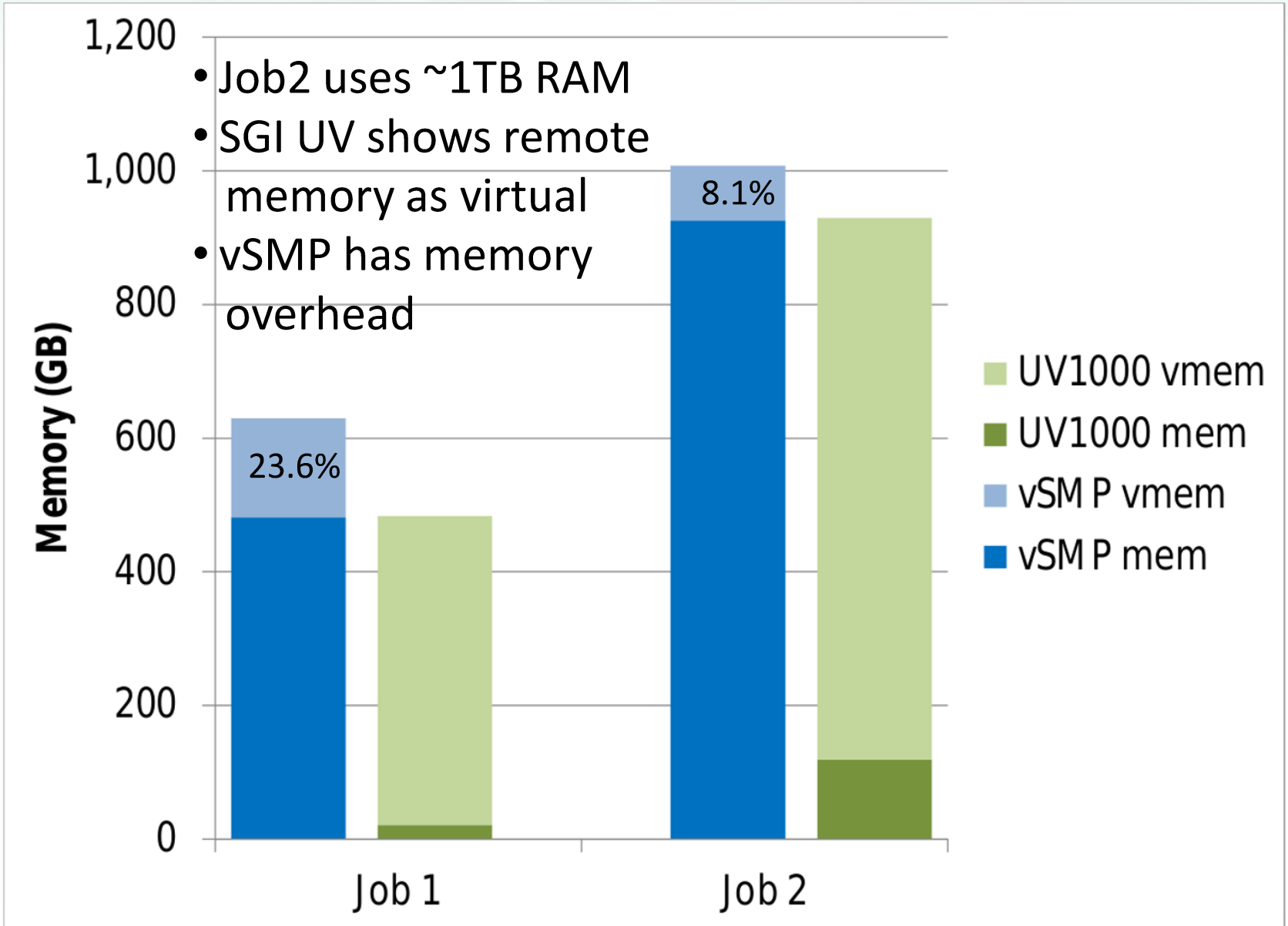
Trinity Wallclock time



Trinity CPU time



Trinity Memory Utilization



Performance Roundup

- Using 192 cores, ScaleMP vSMP performs better than SGI UV 1000 for both SPEC OpenMP and Trinity app
  - Some performance increase simply due to newer CPUs
- Interconnect bandwidth contention can be critical
  - NUMAlink in SGI UV is limited vs QDR InfiniBand in vSMP
  - Research confirms the overhead of remote memory accesses in SMP machines increases as the systems get larger
- SGI UV outperforms vSMP when per-core mem usage is small and the app is cpu bound (cache size differences)
- Both systems capable of handling 1TB+ memory footprints
  - vSMP slightly faster than SGI UV for Trinity, but requires larger memory footprint

SMP Usability

- Overall SMP systems easy to use, especially compared to distributed memory clusters or accelerators
  - SMP systems still require tuning to get optimal performance
  - Pay attention to NUMA efficiencies, bandwidth, etc.
- ScaleMP vSMP has added advantage of hardware re-provisioning between cluster and vSMP
  - Uses commodity servers with InfiniBand, can be rebooted into normal HPC cluster
  - Best suited for systems with variable workloads

Conclusion

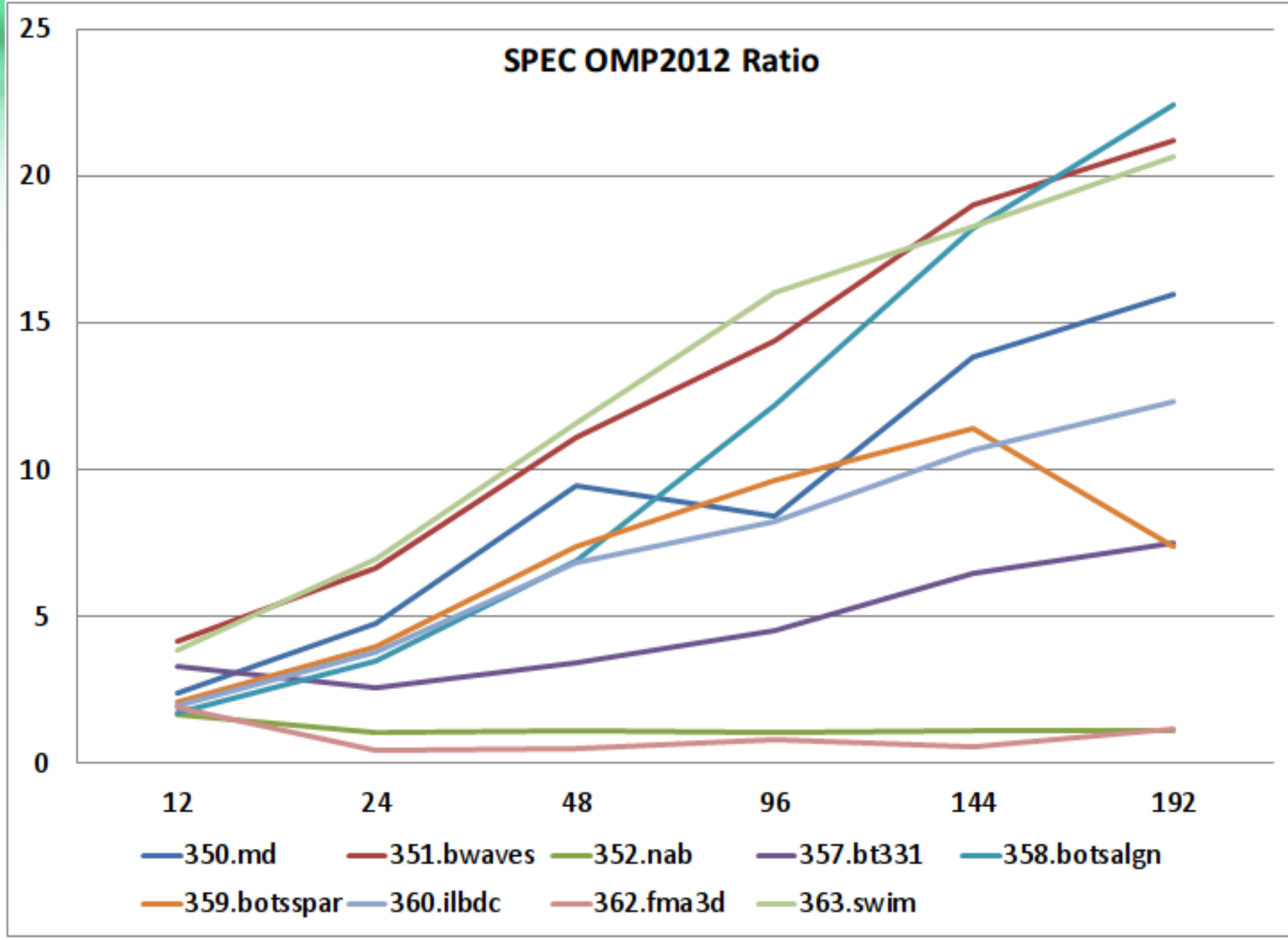
- SMP machines have many uses for big data computation due to their ease of use
- We evaluate two compelling SMP options: ScaleMP vSMP and SGI UV1000 deployed at IU and UofA, respectively
  - Both machines perform well with 192-core OpenMP benchmarks and Trinity de novo assembler
  - vSMP system often faster, due to better hardware
- Due to explosion of data and the complexity of distributed systems, we expect SMP system utilization to increase

Future Work

- Evaluate other SMP systems – IBM POWER8 SMP
- Expand to other Big Data applications
  - Apache Big Data Stack – shared memory/threading vs TCP communication
  - Many Bioinformatics apps
  - Proprietary applications
  - Graph processing
- New ScaleMP deployment at Arizona
  - ~8400core cluster, 9.4TB vSMP license

ajyounge@indiana.edu ajyounge.com

QUESTIONS?


SPEC OMP2012 Ratio
25
20
15
10
5
0
12
24
48
96
144
192
350.md
351.bwaves
352.nab
357.bt331
358.botsalgn
359.botsspar
360.ilbdc
362.fma3d
363.swim