THE HIGH-END VIRTUALIZATION COMPANY

SERVER AGGREGATION – CREATING THE POWER OF ONE

Virtual SMP with vSMP Foundation

Nir Paikowsky

Director of Application Engineering

Agenda

1. INTRODUCTION TO SCALEMP
2. PRODUCT OVERVIEW
3. HOW DOES IT LOOK
4. TYPICAL USE CASES
5. HOW DOES IT WORK
6. PERFORMANCE

1

INTRODUCTION

ScaleMP at a Glance

- Founded in 2003
- Product shipping since 2006
- Sold through Tier-1 and Tier-2 OEMs

Virtualization for high-end computing,
delivering **higher performance**
and lower **Total Cost of Ownership (TCO)**

Aggregation software creates
a **virtual shared-memory multi-processor (SMP)**
from **multiple off-the-shelf x86** servers

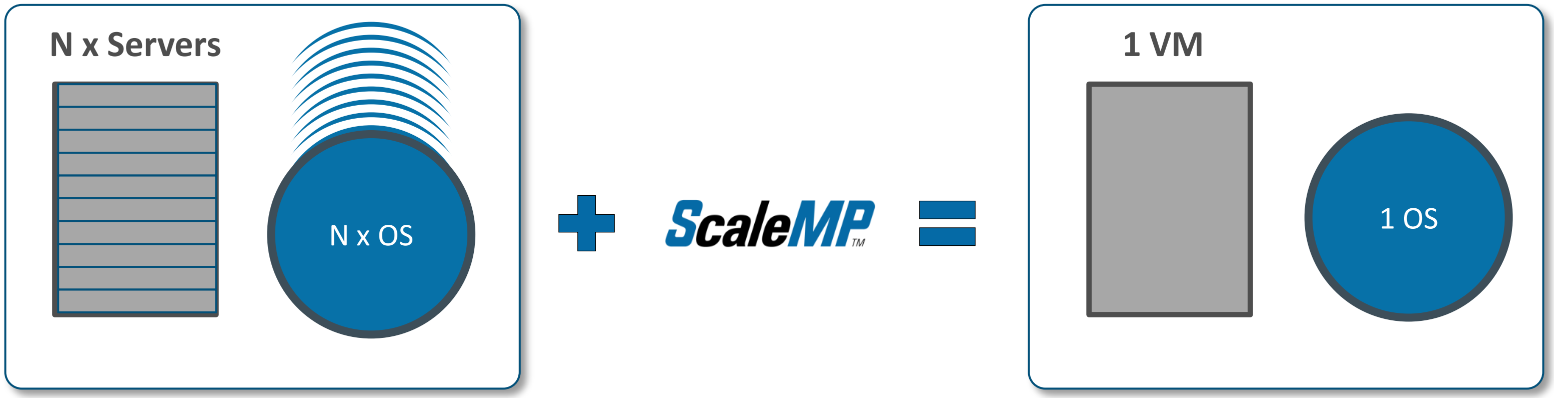
150+ Deployments Worldwide

Commercial

Educational

Federal

What We Do



Virtualization software for aggregating multiple off-the-shelf systems into a single virtual machine, providing improved usability and higher performance

Targeting compute-, memory- and I/O-intensive workloads

2

PRODUCT OVERVIEW

Virtualization Across Different Domains…

…and comparing partitioning and aggregation approaches

Partitioning

Providing a virtual resource that is a **subset** of the physical resource

"Utilization"

Software	Hardware	
Volume Mgmt	Array-based	← Disk Partitioning
Stack-based	Switch-based	← VLANs
Hypervisor / VMM	Mainframe	← Server Virtualization

Aggregation

Providing a virtual resource that is a **concatenation** of several physical resources

"Management, Capability"

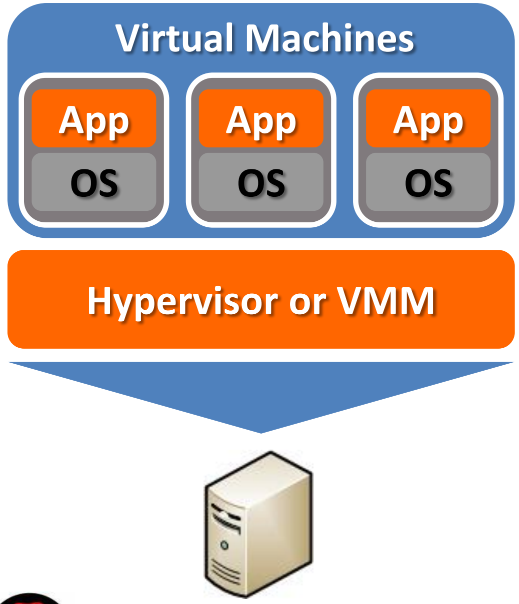
	Hardware	Software
Disk Concatenation →	Array-based	Volume Mgmt
Link Aggregation →	Switch-based	OS-based
Server Aggregation →	SMP, MPP	ScaleMP

Single system only

Approaches to Server Virtualization

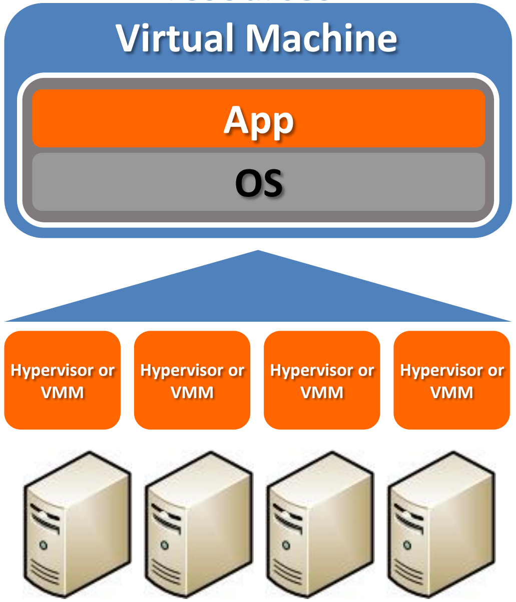
Partitioning

Providing a virtual resource that is a **subset** of the physical resource



Aggregation

Providing a virtual resource that is a **concatenation** of several physical resources



3

HOW DOES IT LOOK

TOP and FREE

```
~
top - 17:52:37 up  3:46,  1 user,  load average: 0.40, 0.31, 0.28
Tasks: 1383 total,   1 running, 1382 sleeping,   0 stopped,   0 zombie
Cpu(s):  0.0%us,  0.0%sy,  0.0%ni,100.0%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Mem:  660813332k total,  5347508k used, 655465824k free,    16096k buffers
Swap:         0k total,        0k used,         0k free,   126996k cached

  PID USER      PR  NI  VIRT  RES  SHR S %CPU %MEM    TIME+  P COMMAND
13431 root      15   0 13676 2112  820 R  2.3  0.0   0:00.20 2 top
    1 root      18   0 10344  680  568 S  0.0  0.0   0:38.39 3 init
    2 root      RT   0     0    0    0 S  0.0  0.0   0:00.02 0 migration/0
    3 root      34  19     0    0    0 S  0.0  0.0   0:00.00 0 ksoftirqd/0
    4 root      RT   0     0    0    0 S  0.0  0.0   0:00.00 0 watchdog/0
    5 root      RT   0     0    0    0 S  0.0  0.0   0:00.02 1 migration/1
    6 root      34  19     0    0    0 S  0.0  0.0   0:00.01 1 ksoftirqd/1
    7 root      RT   0     0    0    0 S  0.0  0.0   0:00.00 1 watchdog/1
    8 root      RT   0     0    0    0 S  0.0  0.0   0:00.00 2 migration/2
    9 root      34  19     0    0    0 S  0.0  0.0   0:00.00 2 ksoftirqd/2
   10 root      RT   0     0    0    0 S  0.0  0.0   0:00.00 2 watchdog/2
   11 root      RT   0     0    0    0 S  0.0  0.0   0:00.00 3 migration/3
   12 root      34  19     0    0    0 S  0.0  0.0   0:00.00 3 ksoftirqd/3
   13 root      RT   0     0    0    0 S  0.0  0.0   0:00.00 3 watchdog/3
   14 root      RT   0     0    0    0 S  0.0  0.0   0:00.00 4 migration/4
   15 root      34  19     0    0    0 S  0.0  0.0   0:00.00 4 ksoftirqd/4
   16 root      RT   0     0    0    0 S  0.0  0.0   0:00.00 4 watchdog/4
   17 root      RT   0     0    0    0 S  0.0  0.0   0:00.06 5 migration/5
   18 root      34  19     0    0    0 S  0.0  0.0   0:00.00 5 ksoftirqd/5
   19 root      RT   0     0    0    0 S  0.0  0.0   0:00.00 5 watchdog/5
   20 root      RT   0     0    0    0 S  0.0  0.0   0:00.00 6 migration/6
   21 root      34  19     0    0    0 S  0.0  0.0   0:00.00 6 ksoftirqd/6
   22 root      RT   0
   23 root      RT   0
   24 root      39  19
```

```
/cygdrive/d/SMP/tmp
[root@dash-1-20 ~]# free -g
             total       used       free     shared    buffers     cached
Mem:           630          5        624          0          0          0
-/+ buffers/cache:          4        625
Swap:            0          0          0
[root@dash-1-20 ~]#
```

/proc/cpuinfo

```
processor       : 0
vendor_id       : GenuineIntel
cpu family      : 6
model           : 26
model name      : Intel(R) Xeon(R) CPU           E5530  @ 2.40GHz
stepping        : 5
cpu MHz         : 2400.083
cache size      : 8192 KB
physical id     : 0
siblings        : 8
core id         : 0
cpu cores       : 8
cpu cores       : 8
apicid          : 0
fpu             : yes
fpu_exception   : yes
cpuid level     : 11
wp              : yes
flags           : fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge m
ant_tsc pni ds_cpl vmx est tm2 ssse3 cx16 xtpr dca popcnt lahf_lm
bogomips        : 4820.91
clflush size    : 64
cache_alignment : 64
address sizes   : 40 bits physical, 48 bits virtual
power management:

processor       : 1
vendor_id       : GenuineIntel
cpu family      : 6
model           : 26
model name      : Intel(R) Xeon(R) CPU           E5530  @ 2.40GHz
stepping        : 5
cpu MHz         : 2400.083
cache size      : 8192 KB
physical id     : 0
siblings        : 8
core id         : 1
cpu cores       : 8
cpu cores       : 8
apicid          : 2
fpu             : yes
fpu_exception   : yes
cpuid level     : 11
wp              : yes
```

```
processor       : 59
processor       : 60
processor       : 61
processor       : 62
processor       : 63
processor       : 64
processor       : 65
processor       : 66
processor       : 67
processor       : 68
processor       : 69
processor       : 70
processor       : 71
processor       : 72
processor       : 73
processor       : 74
processor       : 75
processor       : 76
processor       : 77
processor       : 78
processor       : 79
processor       : 80
processor       : 81
processor       : 82
processor       : 83
processor       : 84
processor       : 85
processor       : 86
processor       : 87
processor       : 88
processor       : 89
processor       : 90
processor       : 91
processor       : 92
processor       : 93
processor       : 94
processor       : 95
processor       : 96
processor       : 97
processor       : 98
processor       : 99
processor       : 100
processor       : 101
processor       : 102
processor       : 103
processor       : 104
processor       : 105
processor       : 106
processor       : 107
processor       : 108
processor       : 109
processor       : 110
processor       : 111
processor       : 112
processor       : 113
processor       : 114
processor       : 115
processor       : 116
processor       : 117
processor       : 118
processor       : 119
processor       : 120
processor       : 121
processor       : 122
processor       : 123
processor       : 124
processor       : 125
processor       : 126
processor       : 127
[root@dash-1-20 ~]# _
```

VSMPVERSION

```
[root@dash-1-20 ~]# vsmpversion
vSMP Version: 2.1.85.29
vSMP Foundation: 2.1.85.29 (Mar 09 2010 16:23:56)

System configuration:
    Boards:        16 (out of 16)
    Processors:    32 x Intel(R) Xeon(R) CPU E5530 @ 2.40GHz (cores: 4)
    Memory:        16 x 49144MB
    Total memory: 786304MB
        vSMP Foundation:      12352MB
        Reserved for cache:  118592MB
        System memory:       655360MB
    Boot device:   [HDD0] ATA ST9250421AS
Serial number:     1000101
System key:        DRN9R-BYEY7-I2EP1-3LJJD-9KU1M-P52
Supported until:   Jul 1 2012
vsmpctl Version:   42.1.0  (Dec 14 2009 16:27:33) HWI Version: 8(4)
[root@dash-1-20 ~]#
```

/sbin/lspci -vt

```
-+-[0000:f0]-+-00.0  Intel Corporation X58 I/O Hub to ESI Port
 |           +-03.0-[0000:f5]--+-00.0  Intel Corporation 82576 Gigabit Network Connection
 |           |                 \-00.1  Intel Corporation 82576 Gigabit Network Connection
 |           +-1f.0  Intel Corporation 82801JIR (ICH10R) LPC Interface Controller
 |           \-1f.2  Intel Corporation 82801JI (ICH10 Family) SATA AHCI Controller
 +-[0000:e0]-+-00.0  Intel Corporation X58 I/O Hub to ESI Port
 |           +-03.0-[0000:e5]--+-00.0  Intel Corporation 82576 Gigabit Network Connection
 |           |                 \-00.1  Intel Corporation 82576 Gigabit Network Connection
 |           +-1f.0  Intel Corporation 82801JIR (ICH10R) LPC Interface Controller
 |           \-1f.2  Intel Corporation 82801JI (ICH10 Family) SATA AHCI Controller
 +-[0000:d0]-+-00.0  Intel Corporation X58 I/O Hub to ESI Port
 |           +-03.0-[0000:d5]--+-00.0  Intel Corporation 82576 Gigabit Network Connection
 |           |                 \-00.1  Intel Corporation 82576 Gigabit Network Connection
 |           +-1f.0  Intel Corporation 82801JIR (ICH10R) LPC Interface Controller
 |           \-1f.2  Intel Corporation 82801JI (ICH10 Family) SATA AHCI Controller
 +-[0000:c0]-+-00.0  Intel Corporation X58 I/O Hub to ESI Port
 |           +-03.0-[0000:c5]--+-00.0  Intel Corporation 82576 Gigabit Network Connection
 |           |                 \-00.1  Intel Corporation 82576 Gigabit Network Connection
 |           +-1f.0  Intel Corporation 82801JIR (ICH10R) LPC Interface Controller
 |           \-1f.2  Intel Corporation 82801JI (ICH10 Family) SATA AHCI Controller
 +-[0000:b0]-+-00.0  Intel Corporation X58 I/O Hub to ESI Port
 |           +-03.0-[0000:b5]--+-00.0  Intel Corporation 82576 Gigabit Network Connection
 |           |                 \-00.1  Intel Corporation 82576 Gigabit Network Connection
 |           +-07.0-[0000:b3]----00.0  LSI Logic / Symbios Logic SAS1068E PCI-Express Fusion-MPT SAS
 |           +-1f.0  Intel Corporation 82801JIR (ICH10R) LPC Interface Controller
 |           \-1f.2  Intel Corporation 82801JI (ICH10 Family) SATA AHCI Controller
 +-[0000:a0]-+-00.0  Intel Corporation X58 I/O Hub to ESI Port
 |           +-03.0-[0000:a5]--+-00.0  Intel Corporation 82576 Gigabit Network Connection
 |           |                 \-00.1  Intel Corporation 82576 Gigabit Network Connection
 |           +-1f.0  Intel Corporation 82801JIR (ICH10R) LPC Interface Controller
 |           \-1f.2  Intel Corporation 82801JI (ICH10 Family) SATA AHCI Controller
 +-[0000:90]-+-00.0  Intel Corporation X58 I/O Hub to ESI Port
 |           +-03.0-[0000:95]--+-00.0  Intel Corporation 82576 Gigabit Network Connection
 |           |                 \-00.1  Intel Corporation 82576 Gigabit Network Connection
 |           +-1f.0  Intel Corporation 82801JIR (ICH10R) LPC Interface Controller
 |           \-1f.2  Intel Corporation 82801JI (ICH10 Family) SATA AHCI Controller
 +-[0000:80]-+-00.0  Intel Corporation X58 I/O Hub to ESI Port
 |           +-03.0-[0000:85]--+-00.0  Intel Corporation 82576 Gigabit Network Connection
 |           |                 \-00.1  Intel Corporation 82576 Gigabit Network Connection
 |           +-1f.0  Intel Corporation 82801JIR (ICH10R) LPC Interface Controller
 |           \-1f.2  Intel Corporation 82801JI (ICH10 Family) SATA AHCI Controller
 +-[0000:70]-+-00.0  Intel Corporation X58 I/O Hub to ESI Port
 |           +-03.0-[0000:75]--+-00.0  Intel Corporation 82576 Gigabit Network Connection
 |           |                 \-00.1  Intel Corporation 82576 Gigabit Network Connection
 |           +-1f.0  Intel Corporation 82801JIR (ICH10R) LPC Interface Controller
 |           \-1f.2  Intel Corporation 82801JI (ICH10 Family) SATA AHCI Controller
 +-[0000:60]-+-00.0  Intel Corporation X58 I/O Hub to ESI Port
 |           +-03.0-[0000:65]--+-00.0  Intel Corporation 82576 Gigabit Network Connection
 |           |                 \-00.1  Intel Corporation 82576 Gigabit Network Connection
 |           +-1f.0  Intel Corporation 82801JIR (ICH10R) LPC Interface Controller
 |           \-1f.2  Intel Corporation 82801JI (ICH10 Family) SATA AHCI Controller
 +-[0000:50]-+-00.0  Intel Corporation X58 I/O Hub to ESI Port
 |           +-03.0-[0000:55]--+-00.0  Intel Corporation 82576 Gigabit Network Connection
 |           |                 \-00.1  Intel Corporation 82576 Gigabit Network Connection
 |           +-1f.0  Intel Corporation 82801JIR (ICH10R) LPC Interface Controller
 |           \-1f.2  Intel Corporation 82801JI (ICH10 Family) SATA AHCI Controller
 +-[0000:40]-+-00.0  Intel Corporation X58 I/O Hub to ESI Port
 |           +-03.0-[0000:45]--+-00.0  Intel Corporation 82576 Gigabit Network Connection
 |           |                 \-00.1  Intel Corporation 82576 Gigabit Network Connection
 |           +-1f.0  Intel Corporation 82801JIR (ICH10R) LPC Interface Controller
 |           \-1f.2  Intel Corporation 82801JI (ICH10 Family) SATA AHCI Controller
 +-[0000:30]-+-00.0  Intel Corporation X58 I/O Hub to ESI Port
 |           +-03.0-[0000:35]--+-00.0  Intel Corporation 82576 Gigabit Network Connection
 |           |                 \-00.1  Intel Corporation 82576 Gigabit Network Connection
 |           +-1f.0  Intel Corporation 82801JIR (ICH10R) LPC Interface Controller
 |           \-1f.2  Intel Corporation 82801JI (ICH10 Family) SATA AHCI Controller
 +-[0000:20]-+-00.0  Intel Corporation X58 I/O Hub to ESI Port
 |           +-03.0-[0000:25]--+-00.0  Intel Corporation 82576 Gigabit Network Connection
 |           |                 \-00.1  Intel Corporation 82576 Gigabit Network Connection
:
```

/proc/partitions

```
[root@dash-1-20 ~]# cat /proc/partitions
major minor  #blocks  name

   8     0  244198584 sda
   8     1   16386268 sda1
   8     2   16386300 sda2
   8     3    2048287 sda3
   8     4          1 sda4
   8     5  209375113 sda5
   8    16  244198584 sdb
   8    17   16386268 sdb1
   8    18   16386300 sdb2
   8    19    2048287 sdb3
   8    20          1 sdb4
   8    21  209375113 sdb5
   8    32   62522712 sdc
   8    48  244198584 sdd
   8    64   62522712 sde
   8    80  244198584 sdf
   8    96   62522712 sdg
   8   112  244198584 sdh
   8   128   62522712 sdi
   8   144  244198584 sdj
   8   160   62522712 sdk
   8   176  244198584 sdl
   8   192   62522712 sdm
   8   208  244198584 sdn
   8   224   62522712 sdo
   8   240  244198584 sdp
  65     0   62522712 sdq
  65    16  244198584 sdr
  65    32   62522712 sds
  65    48  244198584 sdt
  65    64   62522712 sdu
  65    80  244198584 sdv
  65    96   62522712 sdw
  65   112  244198584 sdx
  65   128   62522712 sdy
  65   144  244198584 sdz
  65   160   62522712 sdaa
  65   176  244198584 sdab
  65   192   62522712 sdac
  65   208  244198584 sdad
  65   224   62522712 sdae
  65   240  244198584 sdaf
  66     0  146484375 sdag
  66    16  146484375 sdah
  66    32  146484375 sdai
  66    48  146484375 sdaj
  66    64  146484375 sdak
  66    80  146484375 sdal
  66    96  146484375 sdam
  66   112  146484375 sdan
   9     0   16386176 md0
   9     2  209375040 md2
   9     1   16386176 md1
[root@dash-1-20 ~]#
```

4

TYPICAL USE CASES

Offerings and Customer Benefits

Software Offerings Packaged for Different Deployments		Primary Customer Benefits	
Cluster New management paradigm for small clusters: 4 to 64 nodes	→	Simplified Management	• One system to manage (single OS) • Fewer, larger nodes in large scale deployments • Simplified approach to clustering
Cloud On-the-fly provisioning for compute grids: unlimited scaling	Dynamic →	Improved System Flexibility	• On-the-fly, aggregated VM provisioning and tear down - meeting the needs of dynamic, scale-out environments (e.g. cloud) • Increased cloud utilization: Drive more workloads into cloud infrastructure
SMP High core-count / large memory systems from standard servers	Static →	Cost Effective and High Performance	• Scalable systems built from standard x86 systems – highly cost effective • Compute, memory and I/O scaled independently resulting in top performance • Large memory x86 resource enabling very large workloads

Solving Customer's Problems: Complexity

- Customer Pain:
  - Lack of in-house system administration & expertise:
    - Multi-system management
    - High-speed networks
    - Distributed file-systems

Customer Examples:

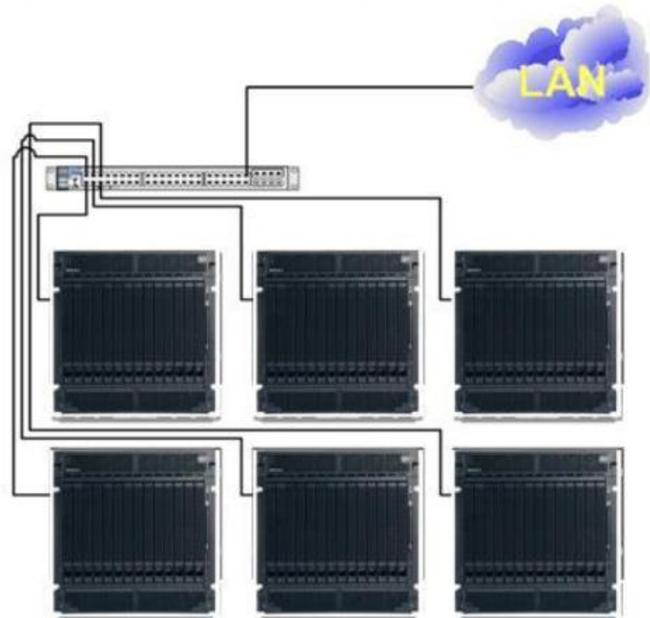
- vSMP Foundation Value:
  - Significantly reduce number of managed systems
  - Reduce the number of tools to operate the environment
  - Enable large/shared memory for performance acceleration and easier programming

- Result:
  - Lower operational costs associated with system management
  - InfiniBand performance without management overhead

Cluster

FAT NODE CLUSTER

- 6-node FAT-NODE compute server, each with 112 cores (Nehalem) and up to 1.3TB RAM using blades and vSMP Foundation
- No Cluster/Parallel File System: Scratch storage using blades' internal drives
- Persistent storage via FTP/NFS over 1GbE or 10GbE



- (6) BladeCenter H " fat-nodes" connected by 10GbE network
- Each " fat-node" provides up to 112 cores (Nehalem) and up to 1.3TB RAM and 8.4TB Storage using blades and vSMP Foundation

Customer Use Cases

FINANCIAL SERVICES

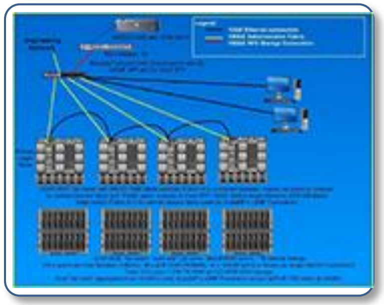
- **Customer:** Hedge Fund
- **Current platform:** Multiple 4-Socket Servers
- **Problems:**
  - A single 4-socket server did not provide enough performance required for customer business targets
  - Co-location at exchanges for a solution comprised of multiple systems is complicate
  - Multiple 4-socket servers required complex decomposition and introduced challenges in transferring data between processes in a short and deterministic time (low latency and small jitters)
    - Ethernet based solution could not provide this / IB solution is too complex to manage and program for
- **Applications:** KX, WOMBAT, Home-grown code
- **Solution:**
  - 16 Intel dual-processor Xeon systems to provide 0.5TB RAM, 32 sockets (128 cores) single virtual system running Linux with vSMP Foundation
  - Alternative considered: IBM P5xx (POWER6). Too expensive and incompatible with x86 application base.
- **Benefits:**
  - **Simpler solution:** Deploy and management of a single system
  - **Simpler programming model:** No need for InfiniBand programming
  - **Better utilization:** Single system reduces resources fragmentation
  - **Performance:** Reduced latency and latency variance

Repeat Customer

SIMPLIFYING INTER-PROCESS COMMUNICATION

Customer Use Cases

ENGINEERING FACULTY



- **Customer:** Engineering Faculty
- **Current platform:** None. Just getting into HPC.
- **Problems:**
  - Compute requirements were growing, as number of users/students was growing
  - No in-house skills to run x86 InfiniBand cluster
  - Limited operational budget to hire additional sys-admin resources
- **Applications:** Commercial code, mostly Fluent and MATLAB
- **Solution:**
  - 4 full blade chassis, each aggregated as a single system with 128 cores and 384 GB RAM and 5 TB of internal storage
  - Total: 64 physical nodes, 512 cores, 20TB storage - running as 4 fat-node cluster
- **Benefits:**
  - **Low OPEX:**
    - No additional IT required for day-to-day operation
    - The need to manage only 4 'Fat-Nodes'
    - Internal storage is embedded in each 'Fat-Node'
  - **Simplicity:** InfiniBand performance without the complexity of managing such a solution

LARGE SCALE DEPLOYMENT WITHOUT THE COMPLEXITY

Solving Customer's Problems: Price & Performance

- Customer Pain:
  - Need faster results with less capital expenditure
  - Purchasing SMP is required but traditional SMP is too expensive
  - Cost of paying for future peak demand upfront

- vSMP Foundation Value:
  - Provide cost effective SMP using x86 commodity hardware
  - Providing more cores at speeds not limited by hardware attributes (virtual solution)
  - Pay as you grow for workloads requiring SMP
  - Supports distributed memory codes as efficiently as a cluster

- Result:
  - Faster run times and ability to run larger problems
  - Lower total cost of ownership than alternatives

Customer Examples:

RWTH AACHEN UNIVERSITY

UF UNIVERSITY of FLORIDA

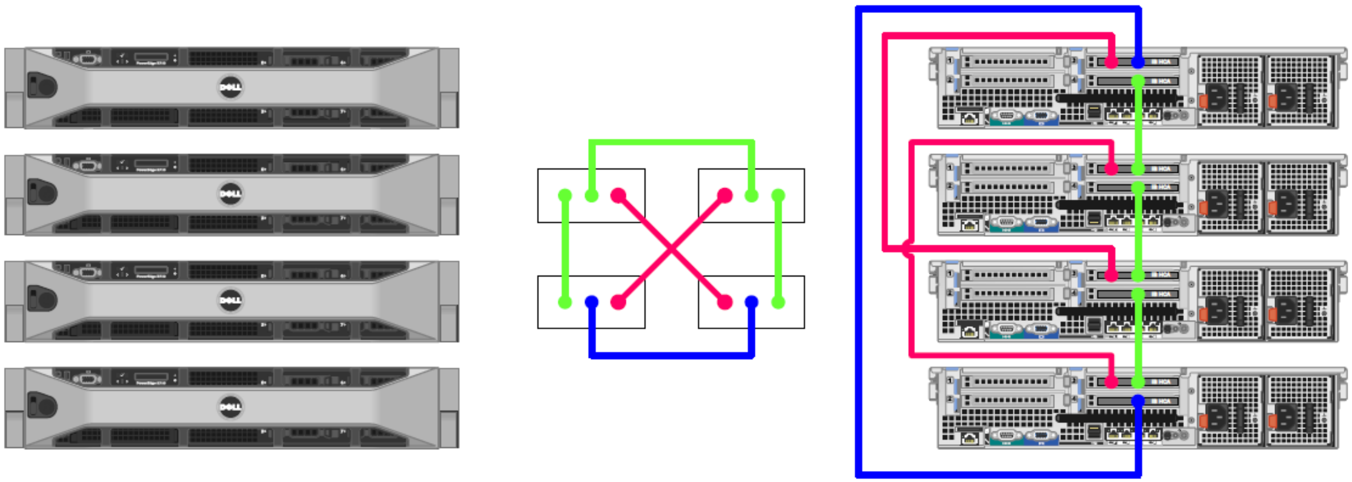
BAMS BARON ADVANCED METEOROLOGICAL SYSTEMS

MGH 1811

SMP 1: Direct Connect 2

vSMP Foundation for SMP - Direct Connect 2

Front **Rear**



4 x Servers:

- 8 x Intel Quad-Core Xeon 5570 2.93GHz
- 576GB RAM (4 x 18 x 8GB DDR3 800MHz)
- 24TB HDD (24 x 3.5" 1TB GB hot-plug SAS hard drives) ***OR*** 1.6TB SSD (32 x 2.5" 50GB SSD)
- 8 x PCIe x4-link Gen 2 expansion slot
- 16 x Ethernet connectors (4 x Integrated 10/100/1000 NIC connectors)

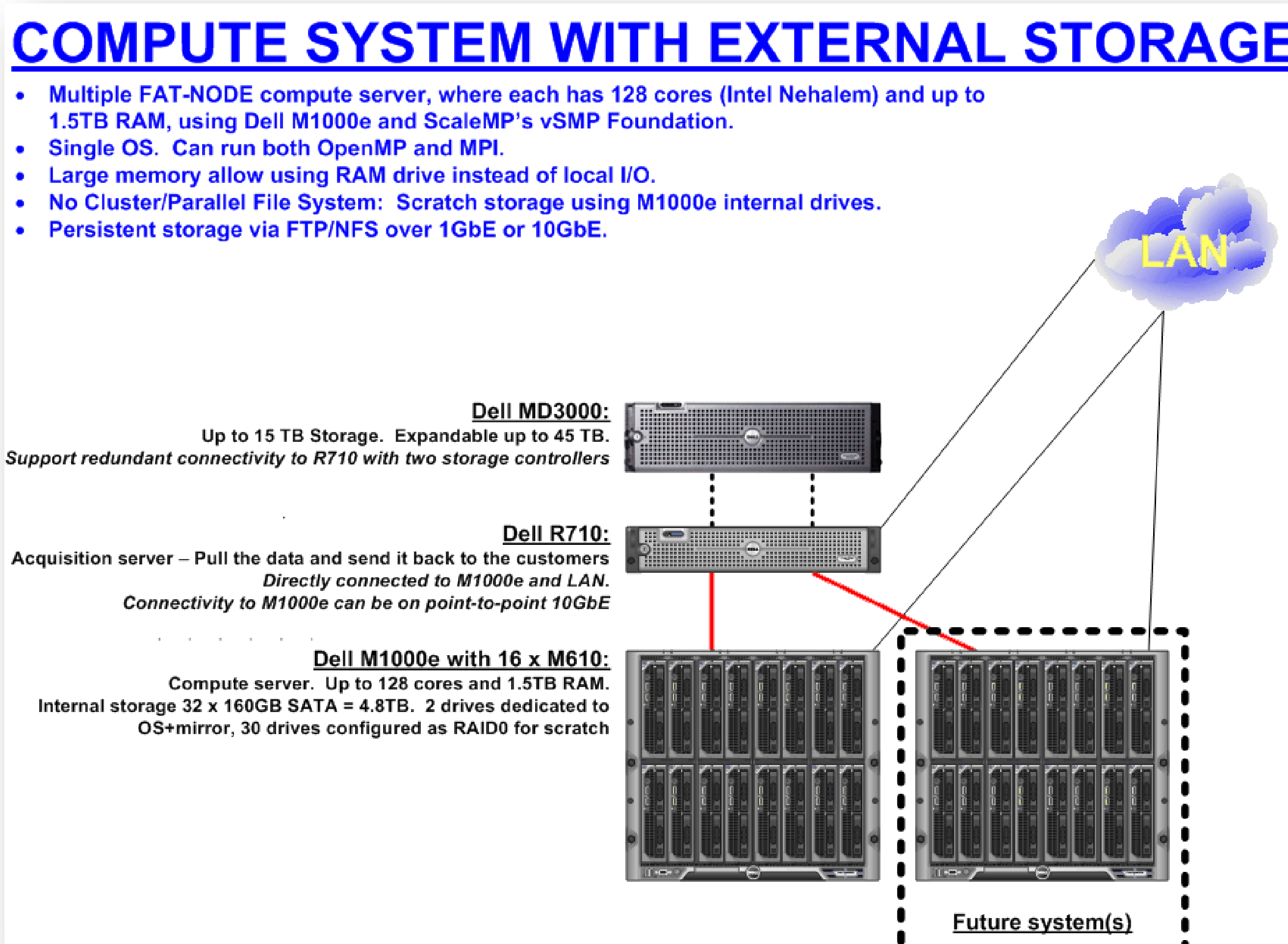
Legend:

Short Cable

Medium Cable

Long Cable

SMP 2


COMPUTE SYSTEM WITH EXTERNAL STORAGE
• Multiple FAT-NODE compute server, where each has 128 cores (Intel Nehalem) and up to 1.5TB RAM, using Dell M1000e and ScaleMP's vSMP Foundation.
• Single OS. Can run both OpenMP and MPI.
• Large memory allow using RAM drive instead of local I/O.
• No Cluster/Parallel File System: Scratch storage using M1000e internal drives.
• Persistent storage via FTP/NFS over 1GbE or 10GbE.
LAN
Dell MD3000:
Up to 15 TB Storage. Expandable up to 45 TB.
Support redundant connectivity to R710 with two storage controllers
Dell R710:
Acquisition server – Pull the data and send it back to the customers
Directly connected to M1000e and LAN.
Connectivity to M1000e can be on point-to-point 10GbE
Dell M1000e with 16 x M610:
Compute server. Up to 128 cores and 1.5TB RAM.
Internal storage 32 x 160GB SATA = 4.8TB. 2 drives dedicated to OS+mirror, 30 drives configured as RAID0 for scratch
Future system(s)

Customer Use Cases

WEATHER FORECASTING SERVICE PROVIDER

- **Customer:** Weather forecasting service provider
- **Current platform:** SGI Altix with 32 cores
- **Problems:**
  - Need to shorten forecast compute times, without limited investment
  - Need to run MPI as well as OpenMP codes
  - System needs to be deployed remotely, and hence needs to be simple to manage
  - Data processing flow is complex and requires transferring large amounts of data between steps
- **Applications:** MM5, WRF, MAWSIP, Home-grown code for data transformation
- **Solution:**
  - 4 Intel Nehalem dual socket blades, total of 8 sockets (32 cores) and 192GB RAM
  - Using high-speed processors and internal storage for best performance
  - Extended to 8 blades, total of 16 sockets (64 cores) and 384GB RAM
- **Benefits:**
  - **Performance:** 2.5 X better performance on same # of cores (32)
  - **Cost:** Faster solution at the cost of annual maintenance of existing platform
  - **Simplicity:** Simple to manage by domain experts (weather forecast scientists)
  - **Dataflow remains within the system, leveraging internal storage**

Repeat Customer

SIMPLE AND FLEXIBLE COST EFFECTIVE SOLUTION

Customer Use Cases

MEDICAL RESEARCH INSTITUTE

- **Customer:** Medical Research Institute
- **Current platform:** HP Superdome System
- **Problems:**
  - Scanned data for a single run is currently over 200GB. Memory requirements are expected to grow significantly with the introduction of full body scan with more sensors
  - Execute high performance image processing on very large MRI scans
  - Would like the ability to use OpenMP and commercial tools for faster development
  - Would like to standardize on x86 architecture due to lower costs and open standards
- **Applications:** Siemens CT processing, MATLAB, BLAS, Home-grown code, …
- **Solution:**
  - 16 Intel dual-processor Xeon systems to provide 1TB RAM, 32 sockets (128 cores) single virtual system running Linux with vSMP Foundation
- **Benefits:**
  - **Performance:** Solution evaluated and found to be faster than alternative systems
  - **Cost:** Significant savings compared to alternative system (order of magnitude)
  - **Versatility:** Also being used for MPI jobs as part of large cluster

LARGE MEMORY FOR MULTI-THREADED PROGRAMMING

Solving Customer's Problems: Inflexibility

- Customer Pain:
  - Mix of distributed and SMP workloads requires dedicated infrastructure per workload
  - Overall system utilization
- vSMP Foundation Value:
  - Homogeneous commodity infrastructure for both distributed and SMP workloads
  - Ability to provision SMP nodes on-demand
  - Reduced OPEX using uniformed hardware infrastructure
- Result:
  - Lower TCO (CAPEX and OPEX)
  - Higher utilization rates optimized for customer workloads

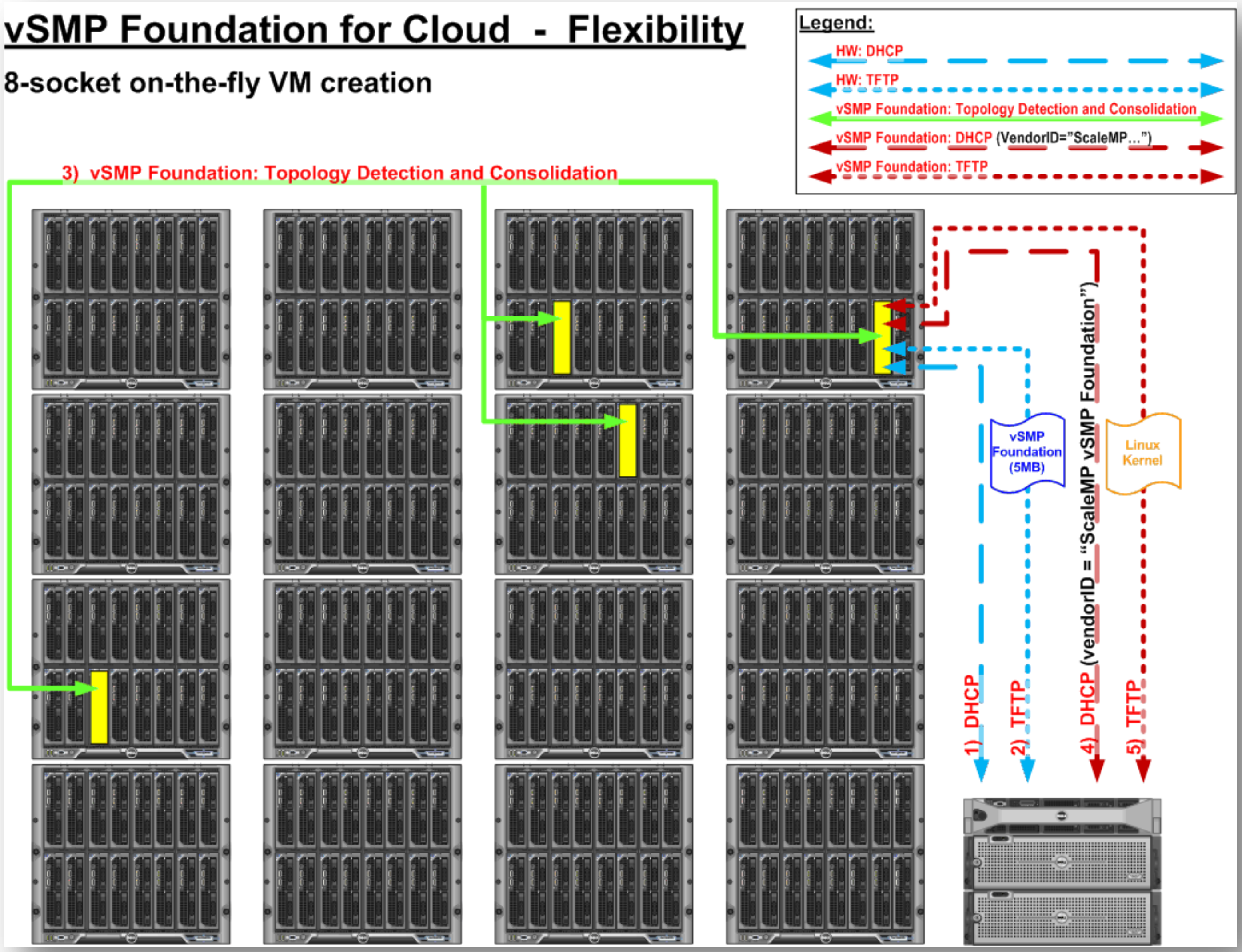
Customer Examples:

r systems
rapid response computing resources

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

vSMP Foundation for Cloud

SINGLE INFRASTRUCTURE – FLEXIBLE RESOURCE PROVISIONING



Customer Use Cases

HOSTED HPC RESOURCE PROVIDER

- **Customer:** Hosted HPC resource provider
- **Current platform:** Clusters and large-memory machines
- **Problems:**
  - Need to run MPI as well as OpenMP (shared memory) codes
  - Large shared memory jobs require dedicated proprietary hardware requiring longer ROI period
  - Low utilization on dedicated shared memory systems
- **Applications:** A variety of commercial codes
- **Solution:**
  - Original: 4 systems, total of 8 sockets (32 cores) and 128GB RAM
  - Solution was extended to 16 nodes – vSMP Foundation for Cloud
- **Benefits:**
  - **Utilization:** Rely on same standard commodity hardware for MPI, large memory, and OpenMP applications
  - **Flexibility:** Being able to provision multiple SMP systems when required, resulting in high utilization and higher income level

Repeat Customer

COST EFFECTIVE FLEXIBLE SOLUTION WITH HIGH UTILIZATION

Customer Use Cases

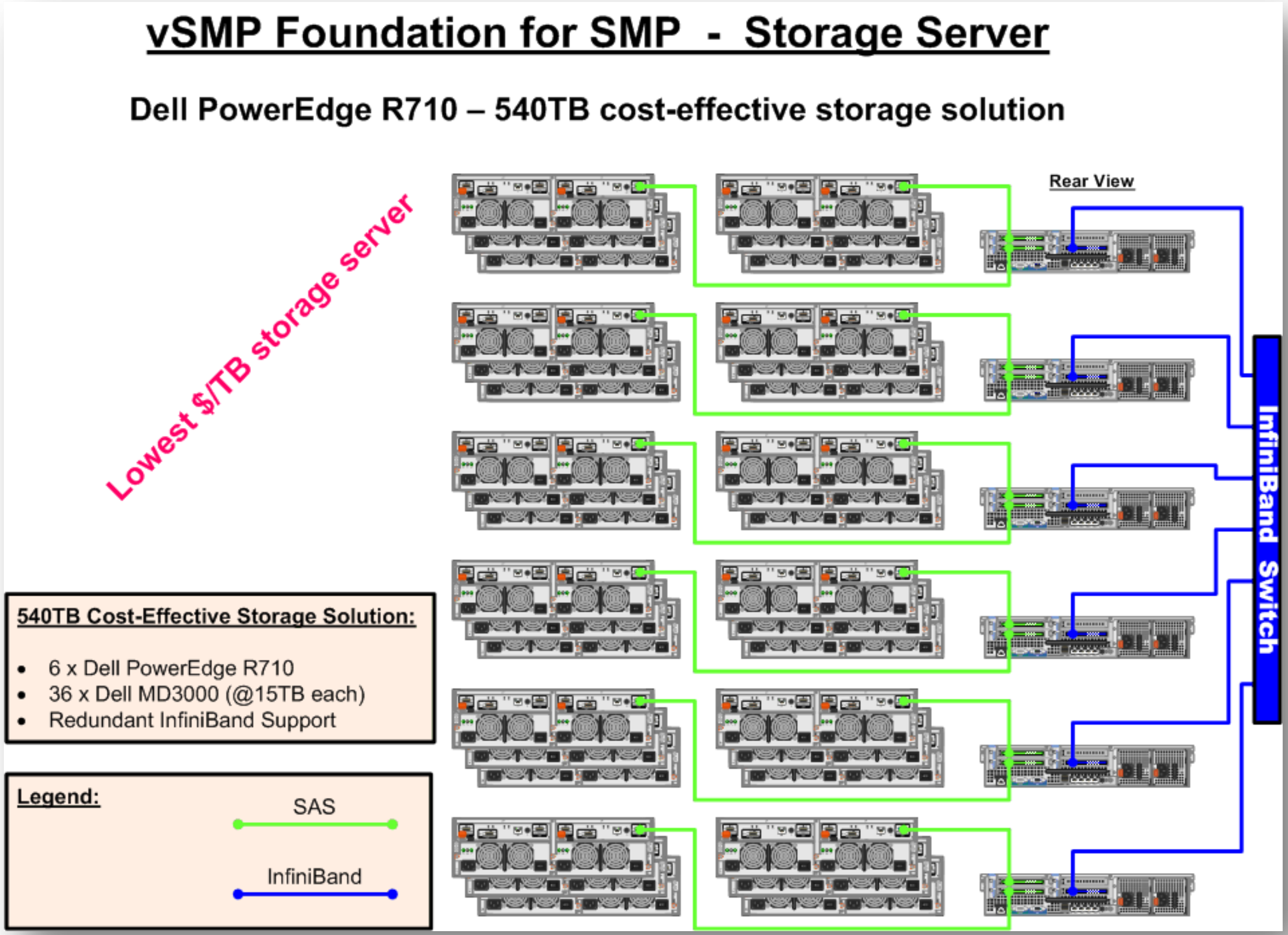
SUPER COMPUTER CENTER

- **Customer:** San Diego Supercomputer Center (SDSC)
- **Current platform:** AMD 8 Socket Systems
- **Problems:**
  - Require an infrastructure for data intensive computing
  - Need large memory system (TBs in size), depending on job need
  - Require the ability to access quickly large amounts of storage
- **Applications:** A variety of data intensive codes (Astronomy, Genomics, Data Mining, etc..)
- **Solution:**
  - Initial Deployment: 4 X ‘Super Nodes’, each with 768GB RAM, 128 Cores, 10TB Internal Storage
  - Complete Deployment (2011): 1,024 servers with vSMP Foundation for Cloud . Could be aggregated up to 32 ‘Super Nodes’ each nodes is 32 servers, resulting in 2TB RAM and 8TB of SSDs
  - On demand allocation using web-request and fast (<10 minutes) provisioning.
- **Benefits:**
  - **Flexibility:** Being able to provision multiple ’Super Nodes’ on various sizes according to need
  - **Performance:** Extremely fast hierarchical memory solution: RAM -> Aggregated RAM -> Aggregated SSDs

ELASTIC VM SOLUTION AIMED FOR DATA INTENTIVE COMPUTING

Storage: vSMP Foundation for SMP

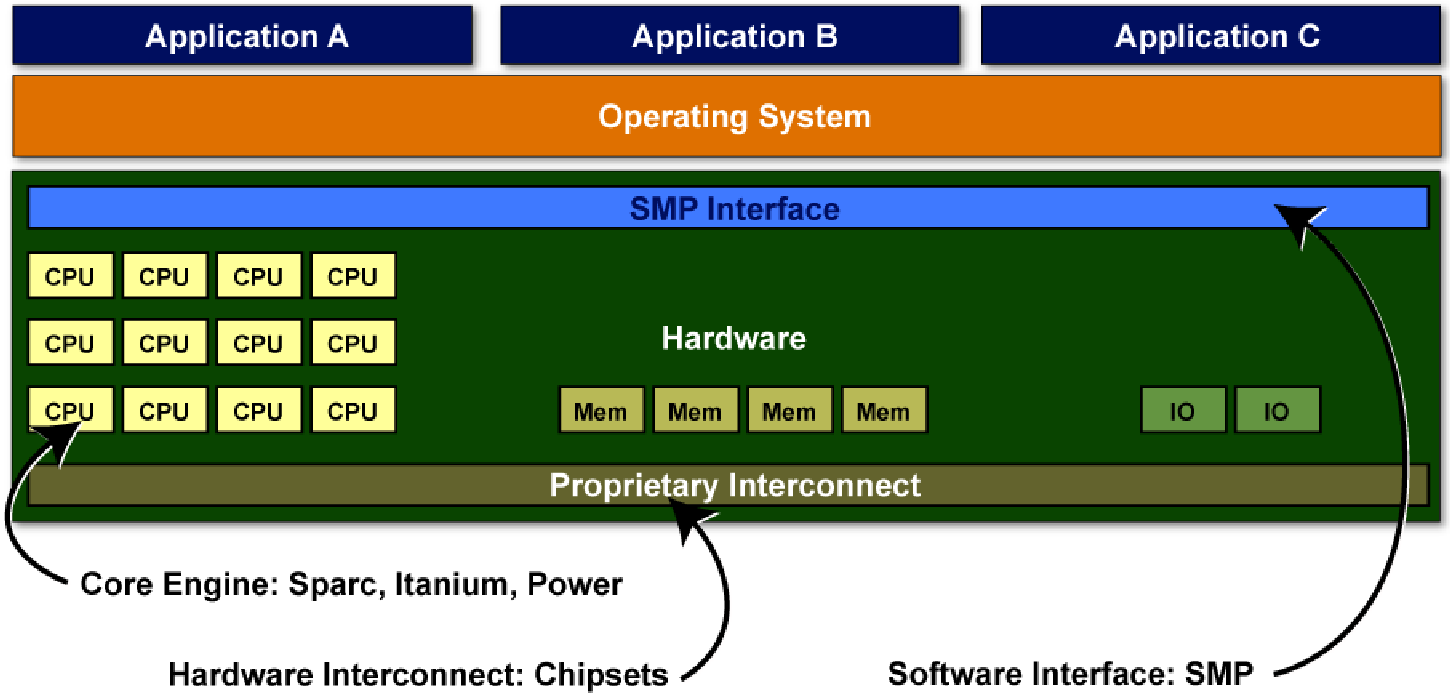
NOT JUST COMPUTE - STORAGE SERVER SOLUTION



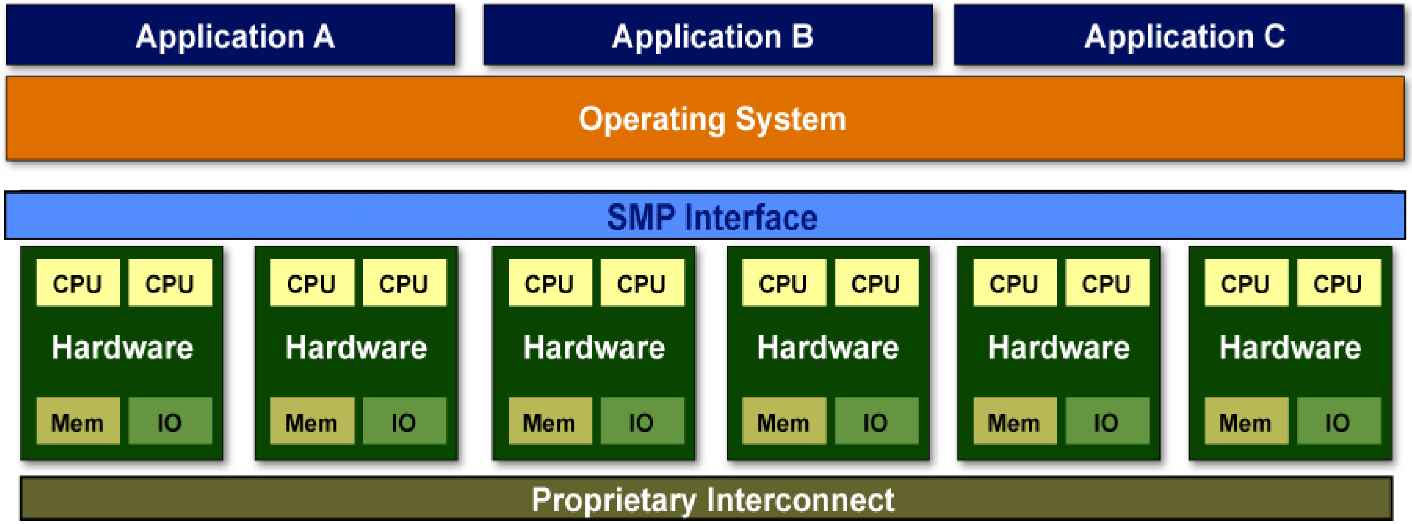
5

HOW DOES IT WORK

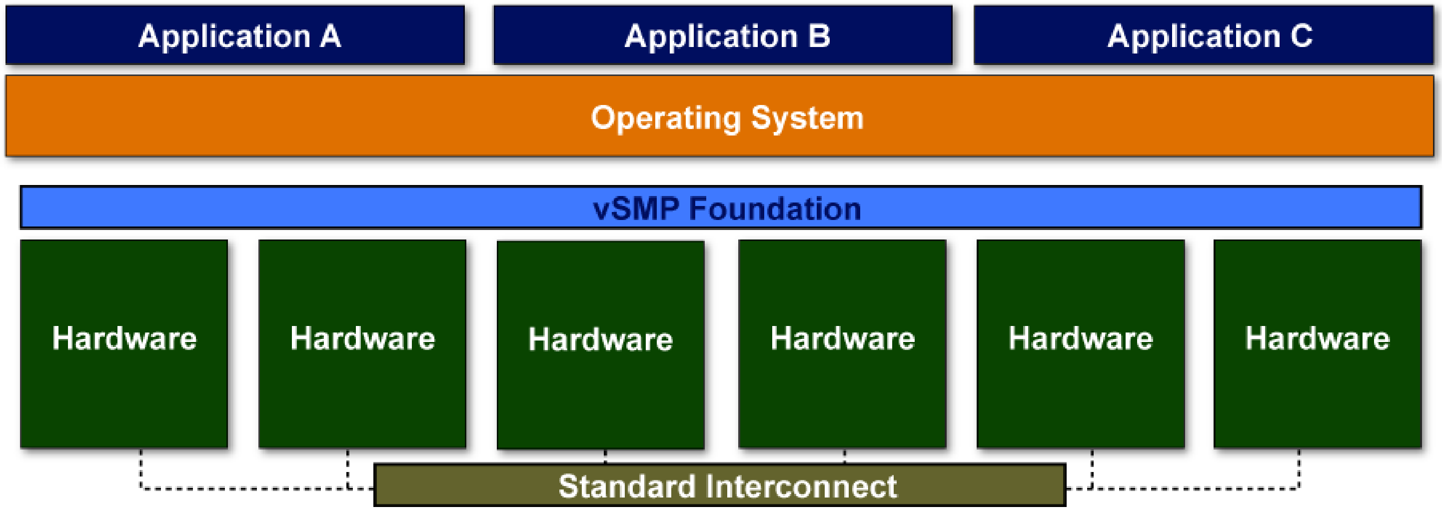
Traditional SMP



Evolving Traditional SMP



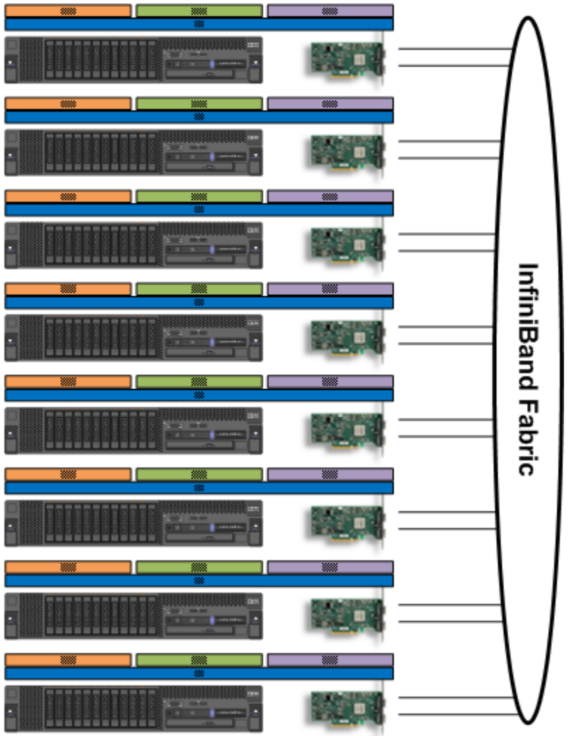
Evolving Traditional SMP

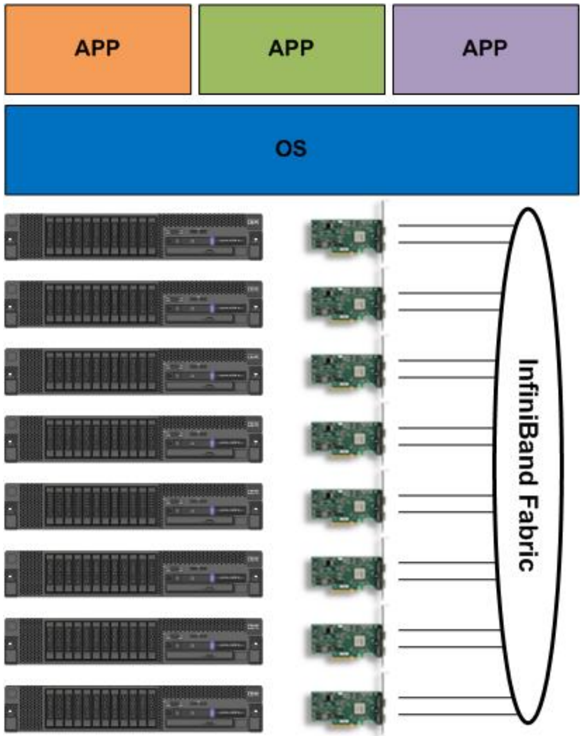


How Does it Work ?

Multiple Computers with Multiple Operating Systems

Multiple Computers with a Single Operating System



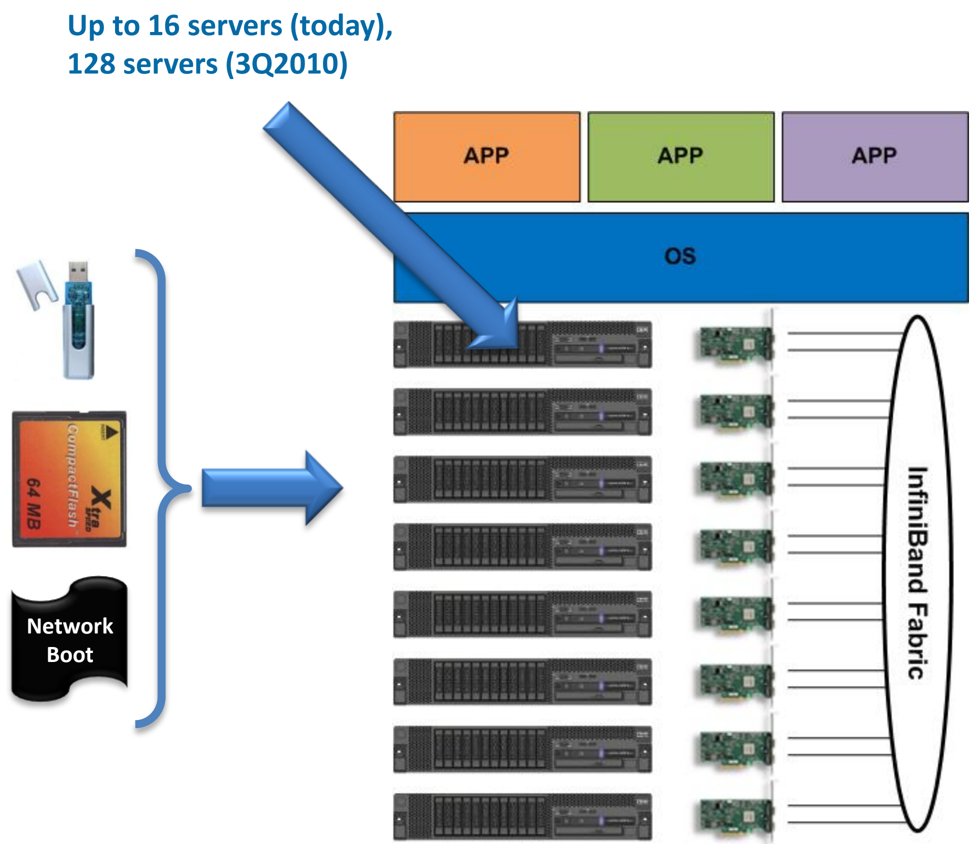


How Does it Work ?

Bare Metal, Distributed Virtual Machine Monitor

- Loaded at boot time
  - Supported boot devices: USB, IDE, CompactFlash or Network Image (PXE)
- Fabric probing and VM setup
- Loading the OS and maintaining I/O and memory coherency

Multiple Computers with a Single Operating System



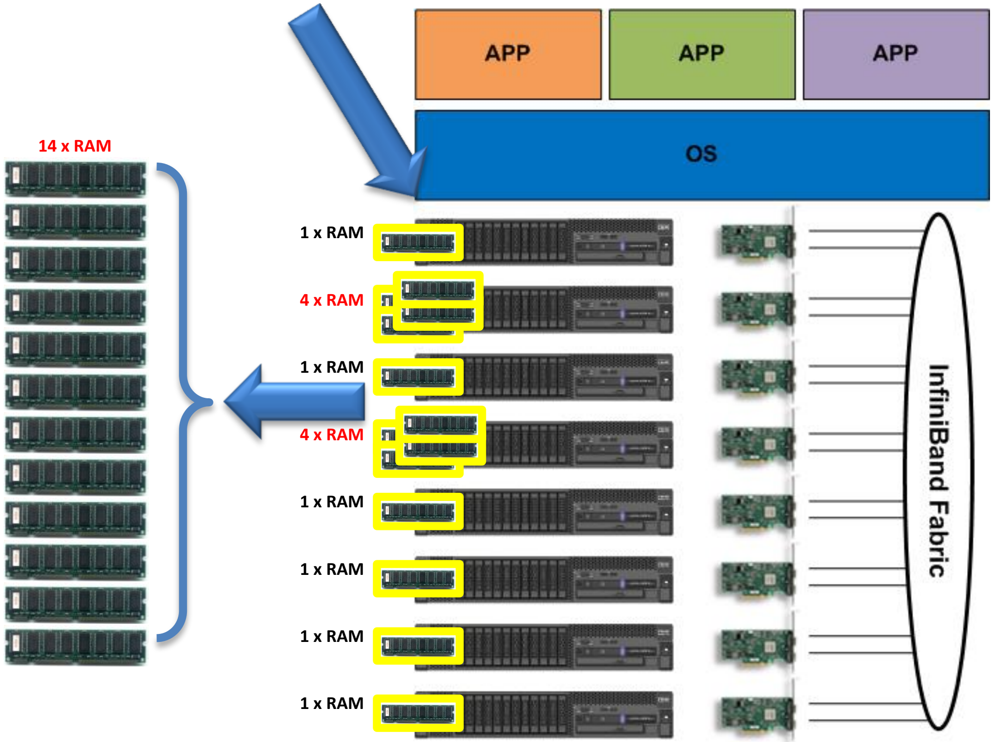
How Does it Work ?

Aggregated System

- Systems configuration can be different
  - Aggregating systems with different boards, I/O configurations, processors speed and memory configuration
  - Only one type of CPU will be presented to the OS
- >10 different coherency mechanisms
- Aggregated hardware I/O compatibility list include devices from Intel, Broadcom, LSI, ATI, Emulex, Adaptec and others

Multiple Computers with a Single Operating System

Up to 4TB aggregated (today), 64TB aggregated (3Q2010)



Behind The Scenes

One System

- Software interception engine creates a uniform execution environment
- vSMP Foundation creates the relevant BIOS environment to present the OS (and the SW stack above it) as single coherent system

Coherent Memory

- vSMP Foundation maintains cache coherency between boards
- Multiple concurrent memory coherency mechanisms, on a per-block basis, based on real-time memory activity access pattern
- Leverage board local-memory for caching

Shared I/O

- vSMP exposes all available I/O resources to the OS in a unified PCI hierarchy
- No need for cluster file systems

Coherent Memory: Efficiency Formula

E = 1 – (A * L)

Efficiency = 1 – (Access * Latency)

ScaleMP's expertise

Fixed, but improving

Coherent Memory: Basics

Trade backplane-latency with redundant RAM

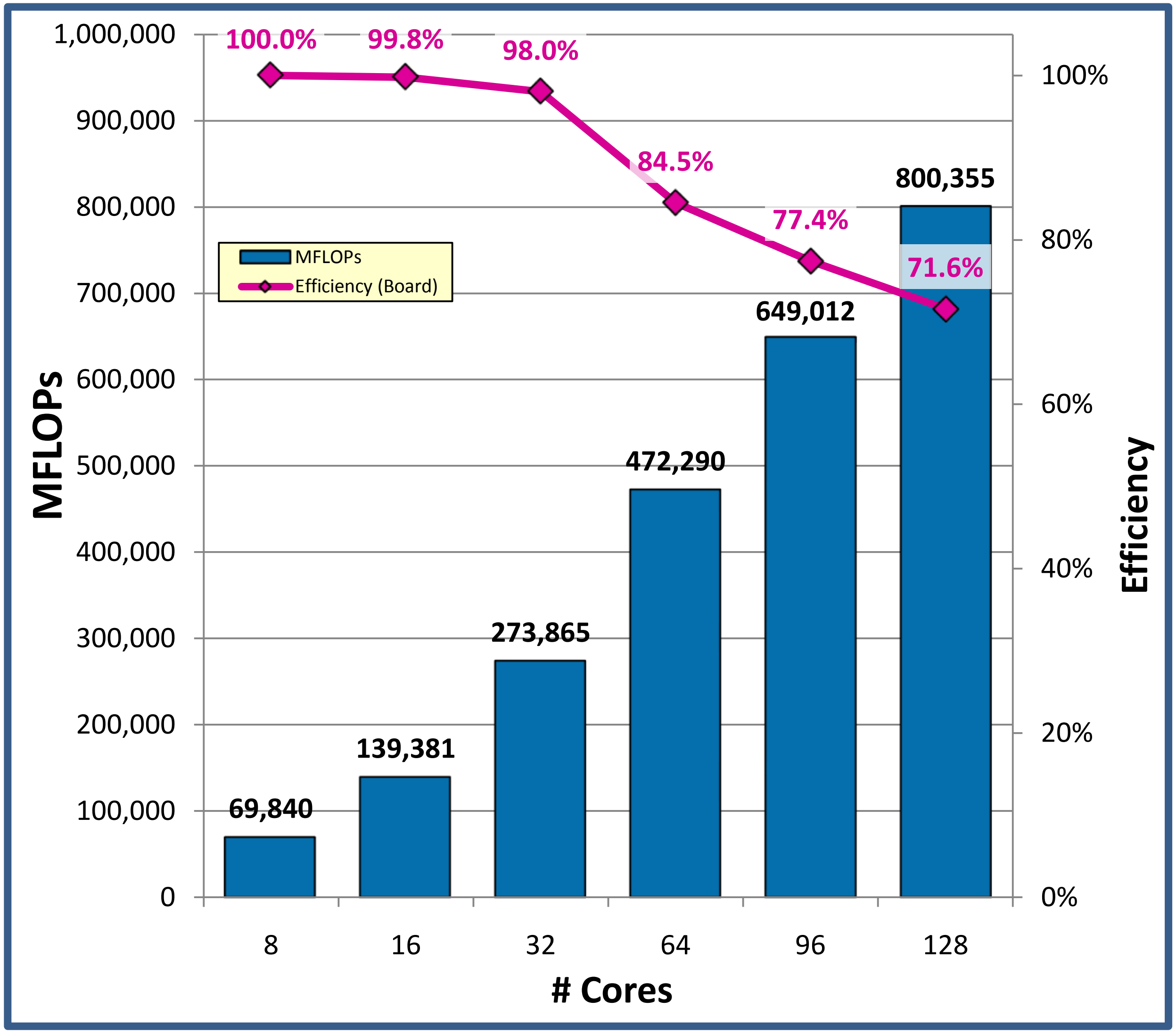
- Hiding backplane latency using software-driven sophisticated and adaptive caching techniques
- Better system economics leveraging PC economies of scale: memory cost vs. propriety backplane/chipset

6

PERFORMANCE EXAMPLES

OPENMP PARALLELIZATION (1)

DGEMM (INTEL MKL)



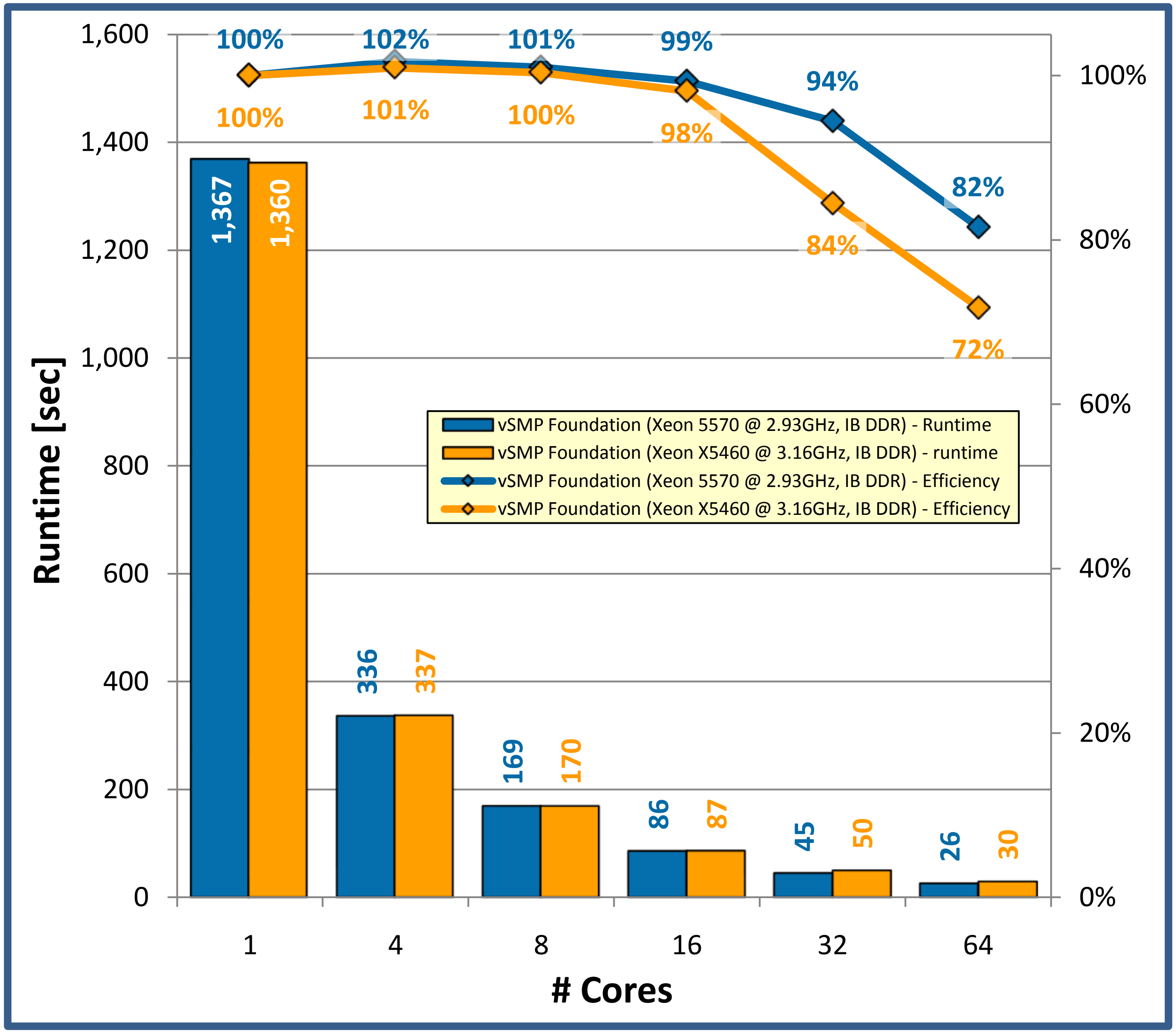
- MKL is Intel's Math Kernel Library, which is using threads for parallelization and is the corner stone for many applications.
- DGEMM is the Matrix Multiply function which is the base for many numerical algorithms. Matrix size used 17,000 X 17,000.
- vSMP Foundation demonstrates over 70% efficiency scaling across 16 boards (128 cores).
- System configuration:
  - /vSMP Foundation (16 nodes) - Data intensive supercomputer system - 128 cores (32 sockets), 768 GB RAM
  - 16 X Dual-socket servers (Intel Xeon E5530 2.40 GHz, 48 GB RAM)

Threaded

OPENMP PARALLELIZATION (2)

Last Update:
9/1/2009

LANCZOS (SMALL)



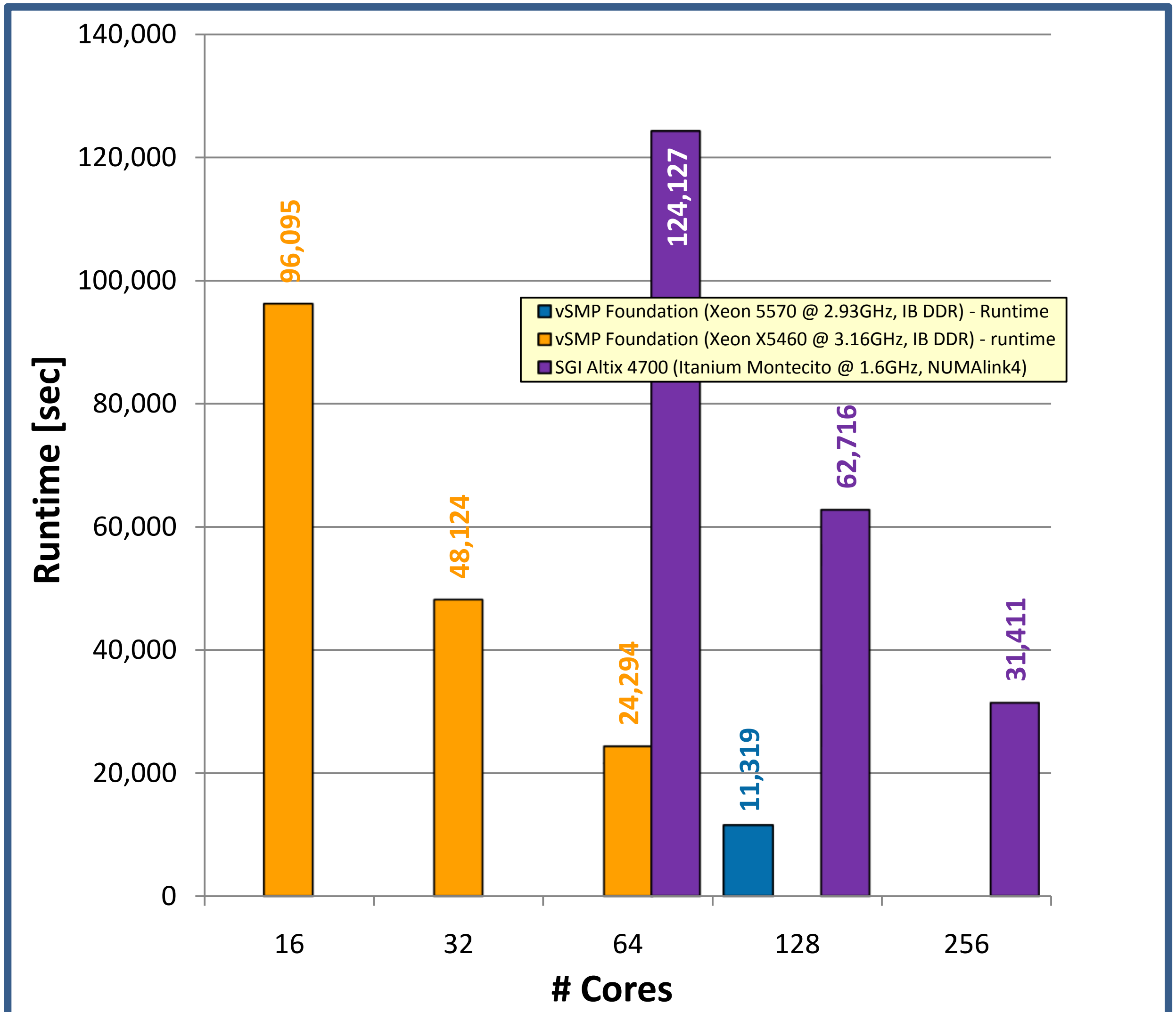
- Customer custom code for calculating eigenvalues leveraging OpenMP for parallelization
- vSMP Foundation demonstrates close to linear scalability using OpenMP:
  - 82% Efficiency with 64 CPUs (Intel Xeon X5570)
  - 72% Efficiency with 64 CPUs (Intel Xeon X5460)

Threaded

OPENMP PARALLELIZATION (3)

Last Update: 9/1/2009

LANCZOS (LARGE)



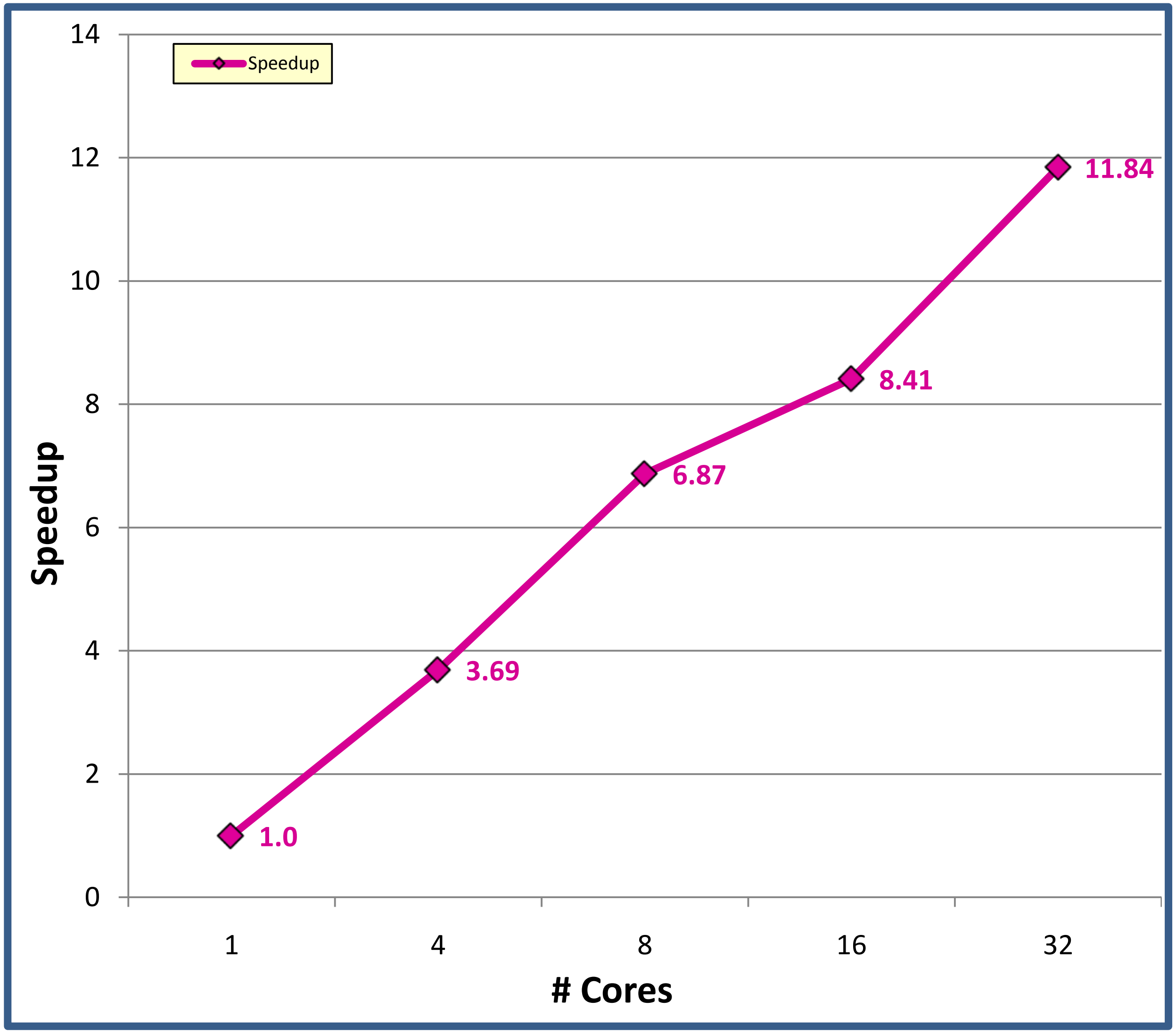
- Customer custom code for calculating eigenvalues leveraging OpenMP for parallelization
- vSMP Foundation demonstrates close to linear scalability using OpenMP
- vSMP Foundation is faster than SGI Altix
  - 11x faster on 64 cores
  - 3x faster with 128 cores compared to Altix with 256 cores

Threaded

GAUSSIAN

Last Update: 9/1/2009

397 BENCHMARK



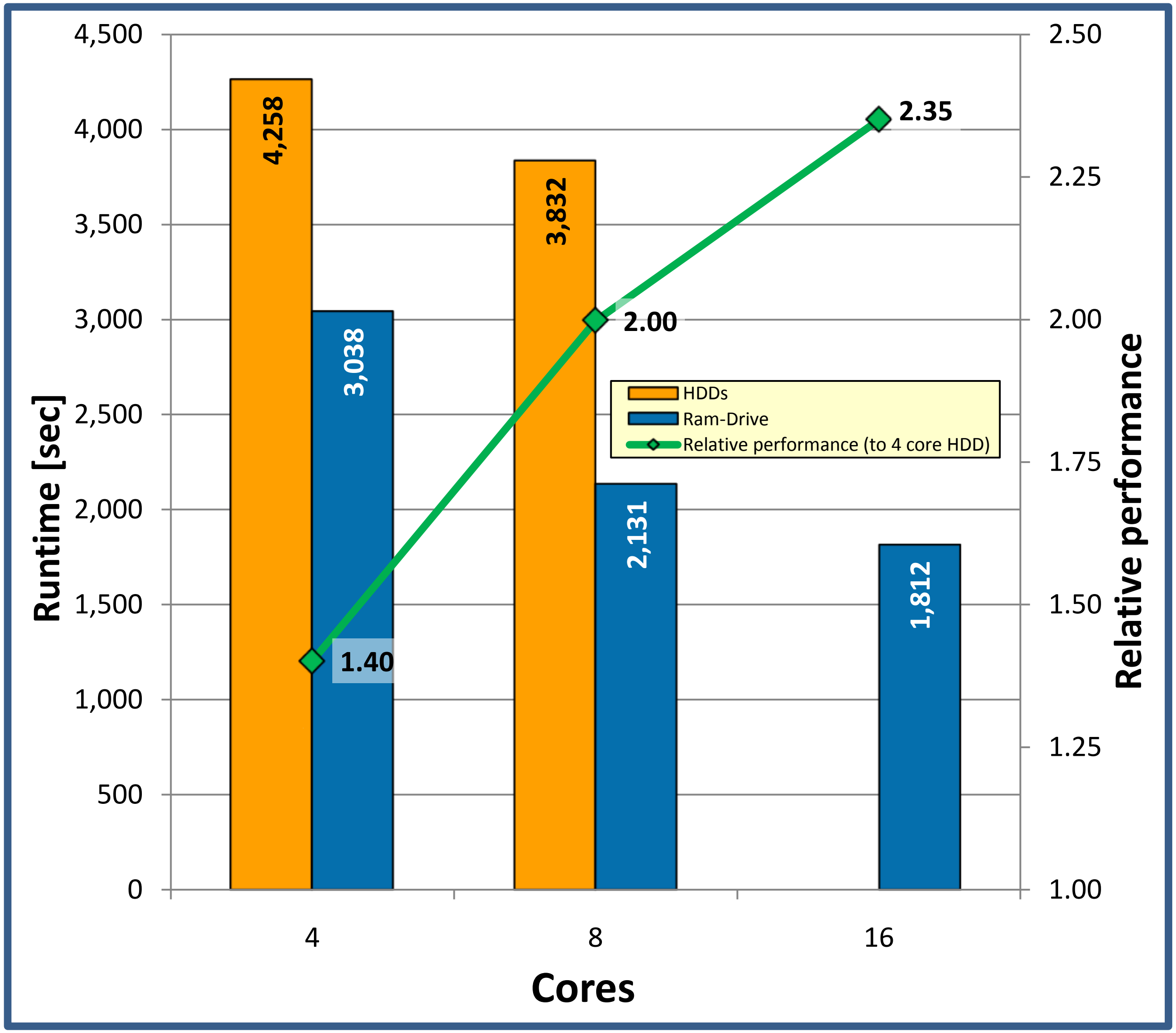
- vSMP Foundation scales up to 32 cores
- System configuration:
  - vSMP Foundation: 16 X Dual-socket servers (Intel Xeon X5570, 2.93 GHz, 48 GB RAM)

Threaded

GAUSSIAN: LARGE MEMORY VS. I/O

Last Update: 9/1/2009

CUSTOMER MOLECULE



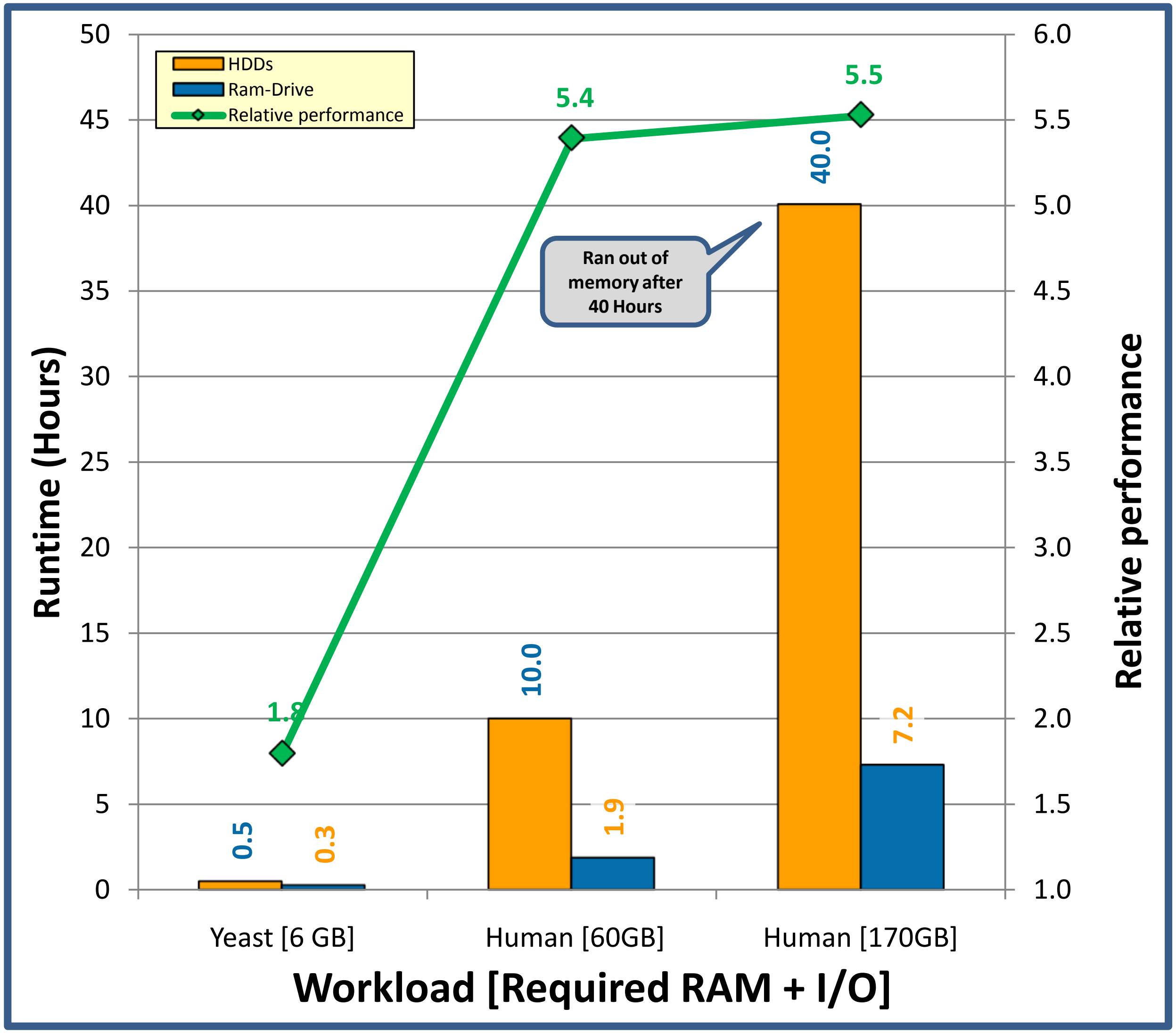
- Comparison of Gaussian workload with limited scalability due to extensive I/O
  - 1 Board system using HDDs
  - Aggregated system using with vSMP Foundation, enabling RAM-drive for I/O
- vSMP Foundation provides improved performance:
  - Scales with aggregated memory for I/O (using RAM-drive)
  - 2.0 X faster compared to HDD performance (8 cores)
  - 2.35 X faster with higher core count (16 cores)
- System configuration:
  - vSMP Foundation: 16 X Dual-socket servers (Intel Xeon X5570, 2.93 GHz, 48 GB RAM)
  - Comparable system: Dual Socket (Intel X5570, 2.93 GHz, 48 GB RAM)

Threaded | Large Memory

454 NEWBLER

Last Update:
9/1/2009

DNA SEQUENCING



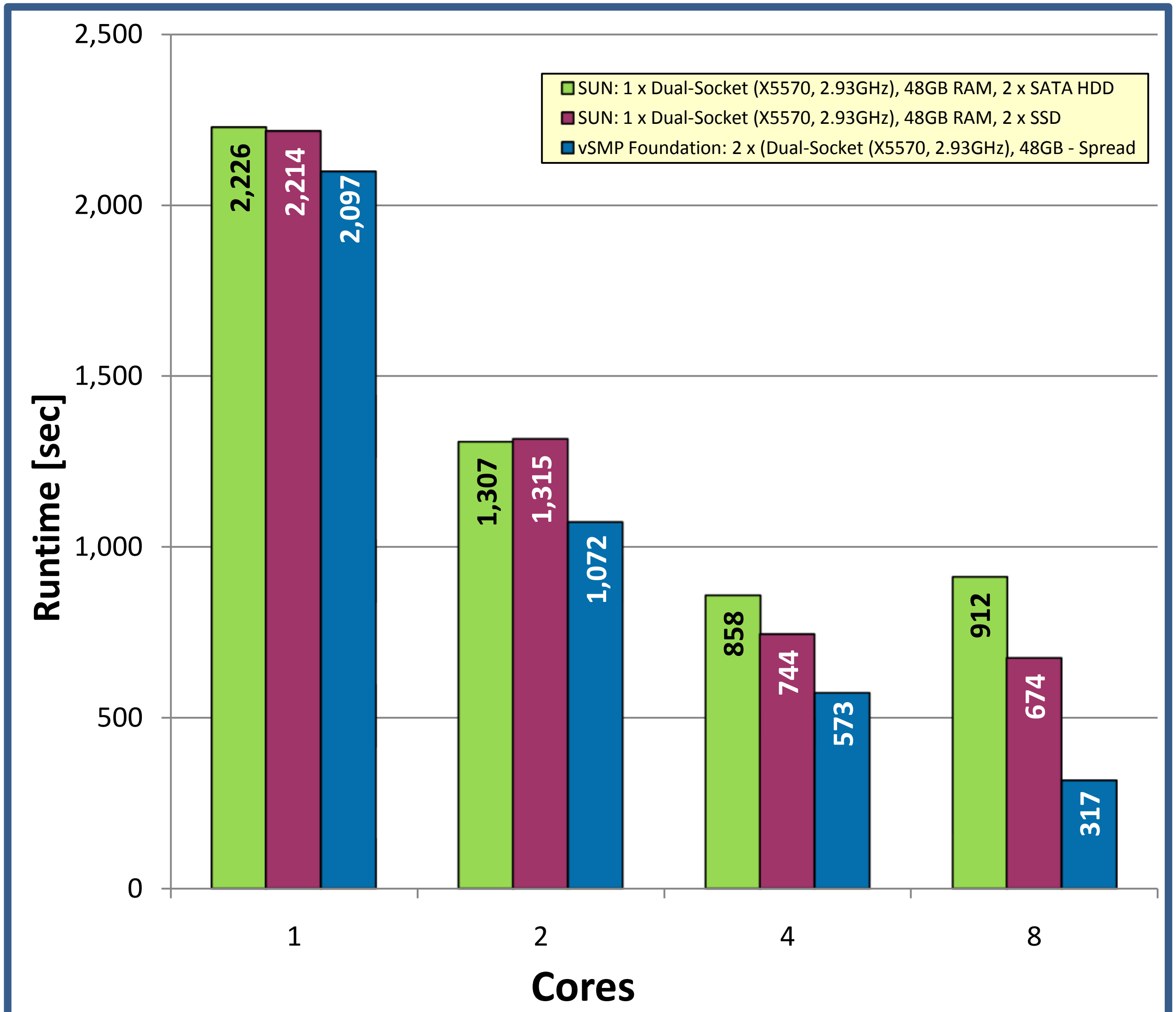
- Performance comparison of DNA Sequencing workload
- Large amounts of memory required, performs extensive I/O and is not using more than 4-8 cores during most of the run
- vSMP Foundation provides improved performance:
  - Running large sequencing inputs which are not feasible otherwise
  - Taking advantage of aggregated RAM for I/O
  - vSMP Foundation achieves 2 to 5 X better performance
- System configuration:
  - Ram-drive: vSMP Foundation: 16 X Dual-socket servers (Intel Xeon X5570, 2.93 GHz, 48 GB RAM)
  - HDD: Dual Socket (X5570 2.93 GHz, 24GB RAM)

Large Memory

MSC NASTRAN (2)

Last Update:
9/1/2009

LARGE MEMORY VS. I/O: XL0TDF1



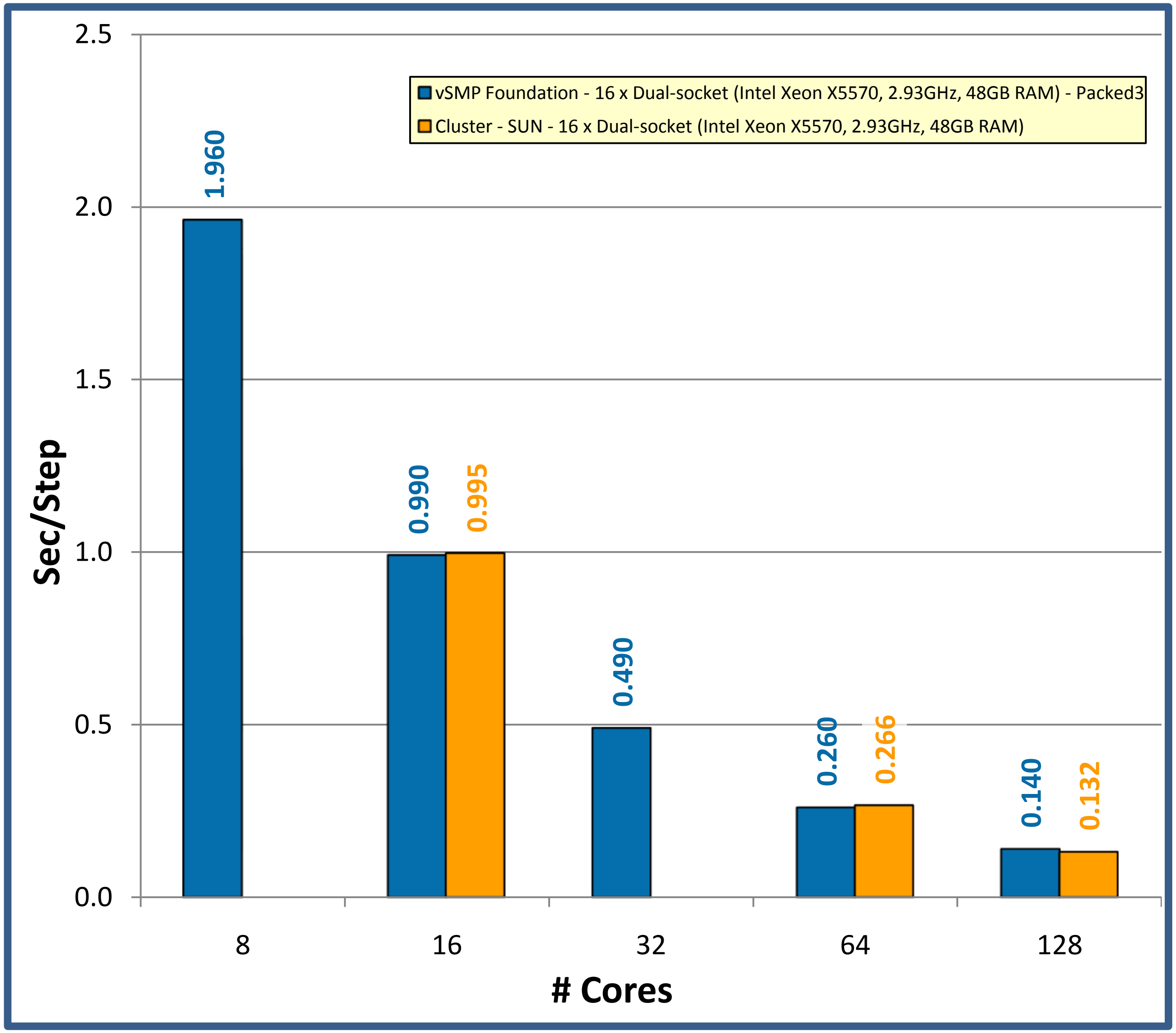
- Performance comparison of:
  - Locally attached HDDs
  - Locally attached SSDs
  - Aggregated memory and CPUs of 2 systems with vSMP Foundation
    - Compute utilized 4 cores on each system, not affecting the application (NASTRAN) license cost
- vSMP Foundation provide significant performance gains:
  - 3 X faster compared to HDDs
  - 2 X compared to SSDs

Throughput / MPI | | Large Memory

NAMD

Last Update:
1/4/2010

STMV MOLECULE



- Performance comparison of:
  - SUN Cluster
  - vSMP Foundation: 16 nodes
- vSMP Foundation provide similar performance to a cluster for MPI based applications
- System configuration:
  - vSMP Foundation: 16 X Dual-socket servers (Intel Xeon X5570, 2.93 GHz, 48 GB RAM)
  - SUN: Sun Blade X6275 12 X 2 Dual-socket (Intel Xeon X5570, 2.93 GHz, 24GB RAM)
- Source: SUN Blog -
  - http://blogs.sun.com/BestPerf/entry/sun_blade_6048_and_sun

Throughput / MPI

7

MORE INFORMATION

More Information

http://www.scalemp.com

http://www.scalemp.com/performance

info@scalemp.com

THE HIGH-END VIRTUALIZATION COMPANY

SERVER AGGREGATION – CREATING THE POWER OF ONE

Nir Paikowsky
Director of Application Engineering

nir@ScaleMP.com, +1 (650) 283 2110