FAST-OS: Petascale Single System Image

Presented by

Jeffrey S. Vetter (PI)
Nikhil Bhatia, Collin McCurdy, Phil Roth, Weikuan Yu

Future Technologies Group
Computer Science and Mathematics Division
PetaScale single system image

• What?
  – Make a collection of standard hardware look like one big machine, in as many ways as feasible

• Why?
  – Provide a single solution to address all forms of clustering
  – Simultaneously address availability, scalability, manageability, and usability

• Components
  – Virtual cluster using contemporary hypervisor technology
  – Performance and scalability of parallel shared root file system
  – Paging behaviors of applications, i.e., reducing TLB misses
  – Reducing “noise” from operating systems at scale
Virtual clusters using hypervisors

- Physical nodes run Hypervisor in the privileged mode.
- Virtual machines run on top of the hypervisor.
- Virtual machines are compute nodes hosting user-level distributed memory (e.g., MPI) tasks.
- Virtualization enables dynamic cluster management via live migration of compute nodes.
- Capabilities for dynamic load balancing and cluster management ensuring high availability.
Virtual cluster management

- Single system view of the cluster
- Cluster performance diagnosis
- Dynamic node addition and removal ensuring high availability
- Dynamic load balancing across the entire cluster
- Innovative load balancing schemes based on distributed algorithms
Virunga—virtual cluster manager
Paging behavior of DOE applications

Goal: Understand behavior of applications in large-scale systems composed of commodity processors

1. With performance counters and simulation:
   • Show that the HPCC benchmarks, meant to characterize the memory behavior of HPC applications, do not exhibit the same behavior in the presence of paging hardware as scientific applications of interest to the Office of Science
   • Offer insight into why that is the case

2. Use memory system simulation to determine whether large page performance will improve with the next generation of Opteron processors
Experimental results (from performance counters)

- Performance: trends for large vs small nearly opposite
- TLBs: significantly different miss rates
Re-use distances (simulated)

Patterns are clearly significantly different…
Paging conclusions

- HPCC benchmarks are not representative of paging behavior of typical DOE applications.
  - Applications access many more arrays concurrently.

- Simulation results (not shown) indicate the following:
  - New paging hardware in next-generation Opteron processors will improve large page performance.
    - Performance near that with paging turned off.
  - However, simulations also indicate that the mere presence of a TLB is likely degrading performance, whether paging hardware is on or off.

- More research into implications of paging, and of commodity processors in general, on performance of scientific applications is required.
Parallel root file system

• **Goals of study**
  – Use a parallel file system for implementation of shared root environment
  – Evaluate performance of parallel root file system
  – Evaluate the benefits of high-speed interconnects
  – Understand root I/O access pattern and potential scaling limits

• **Current status**
  – RootFS implemented using NFS, PVFS-2, Lustre, and GFS
  – RootFS distributed using ramdisk via etherboot
    • Modified mkinitrd program locally
    • Modified to init scripts to mount root at boot time
  – Evaluation with parallel benchmarks (IOR, b_eff_io, NPB I/O)
  – Evaluation with emulated loads of I/O accesses for RootFS
  – Evaluation of high-speed interconnects for Lustre-based RootFS
Example results: Parallel benchmarks
IOR read/write throughput

NFS Write 256MB Block
PVFS2 Write 256MB Block
Lustre Write 256MB Block

NFS Read 256MB Block
PVFS2 Read 256MB Block
Lustre Read 256MB Block
Performance with synthetic I/O accesses

Startup with different images

- Execution Time (sec)
- Number of Processors
- 16K
- 16M
- 1G

Time tar jcf linux-2.6.17.13
- Time (sec)
- Number of Nodes
- CPU Utilization (%)

Time tar jxf linux-2.6.17.13.tar.bz2
- Time (sec)
- Number of Nodes
- CPU Utilization (%)

Time diff mpich2
- Time (sec)
- Number of Nodes
- CPU Utilization (%)

Time cvs co mpich2
- Time (sec)
- Number of Nodes
- CPU Utilization (%)

- Lustre-TCP-Time
- Lustre-TCP-CPU
- Lustre-IB-Time
- Lustre-IB-CPU
Contacts

Jeffrey S. Vetter
Principle Investigator
Future Technologies Group
Computer Science and Mathematics Division
(865) 356-1649
vetter@ornl.gov

Nikhil Bhatia
(865) 241-1535
bhatia@ornl.gov

Collin McCurdy
(865) 241-6433
cmcccurdy@ornl.gov

Phil Roth
(865) 241-1543
rothpc@ornl.gov

Weikuan Yu
(865) 574-7990
wyu@ornl.gov