BG/L I/O and Filesystems

Todd Inglett
IBM Rochester

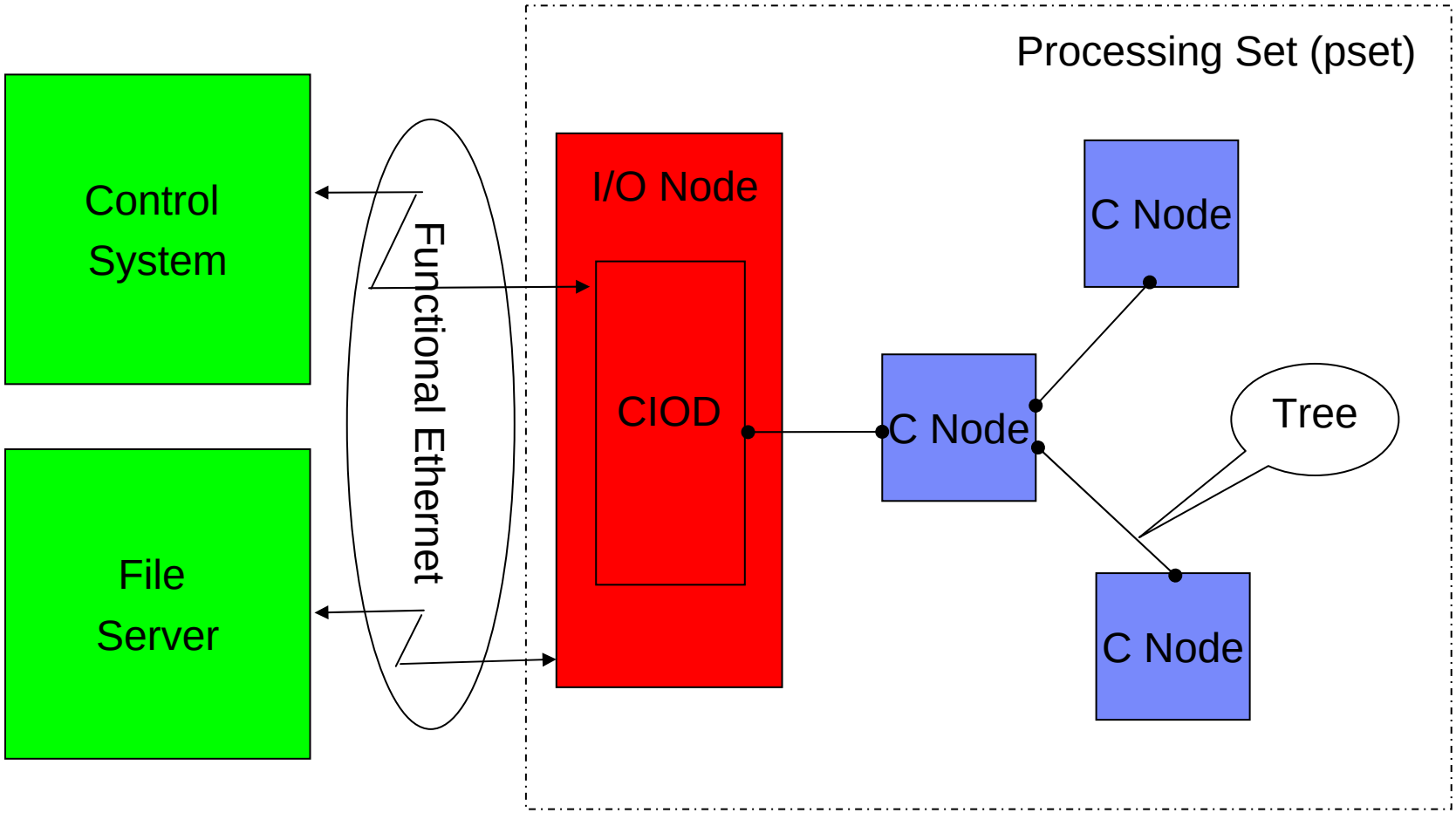
BG/L I/O and Filesystems Agenda

- I/O quick overview
- I/O Tuning and NFS
- GPFS
- Sockets
- Additional function in Release 2 and Release 3

I/O and Filesystems Overview

- CNK redirects I/O syscalls to I/O node
- I/O node is Linux based
  - 2.6.5 kernel based on SLES patches
  - Embedded Linux – no persistent storage or swap
- ciod provides service to CNK
  - Control and I/O Daemon for multiple compute nodes
  - Accepts control connection from service node
  - Loads and starts application
  - Services I/O syscalls for CNK
  - Handles the debugger
- Goal is to remain simple and let Linux handle the filesystem and sockets

I/O Node with ciod



Embedded Linux

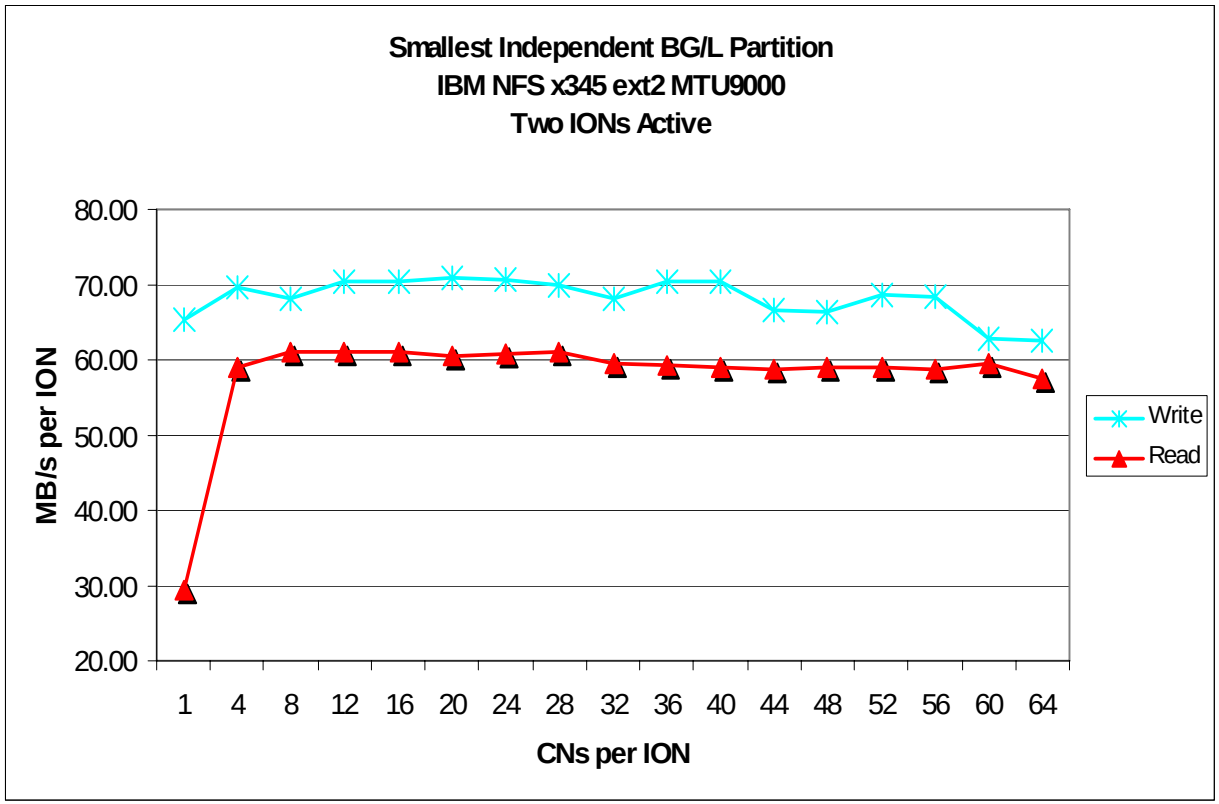
- I/O Node has no persistent storage
- Boot process starts with a ramdisk and uses NFS for startup
- The Startup NFS server is typically the service node but can be customized on a per-midplane basis for load balancing
  - See ionode.README for more details
- Boot runs standard /etc/rc.d/rc3.d/Snn* scripts
  - For site customization, scripts in the installed rpm are numerically merged with site scripts in /bgl/dist/etc/rc.d/rc3.d
  - Advanced customization is allowed by using the installed build-ramdisk script to produce a ramdisk with additional scripts/commands available before NFS is mounted (or to reduce load on the NFS server)
  - See ionode.README for more details
- This boot method is a tradeoff between ramdisk size (and load time) vs. load on NFS server(s)

I/O Tuning

- There are many aspects to tuning the I/O subsystem
  - Boot tuning
  - Network interface tuning
  - Using subnets
  - Using multiple NFS servers
  - Tuning a parallel filesystem
- MTU 1500 -> 9000
- Server settings (sysctl net.core and net.ipv4)
- CIOD_RDWR_BUFFER_SIZE = 524288 (or greater)
- When reading, use 4 or more compute nodes per IONode
- NFS mount options: rsize=wsize=32768,tcp,async
- Try MPI-IO

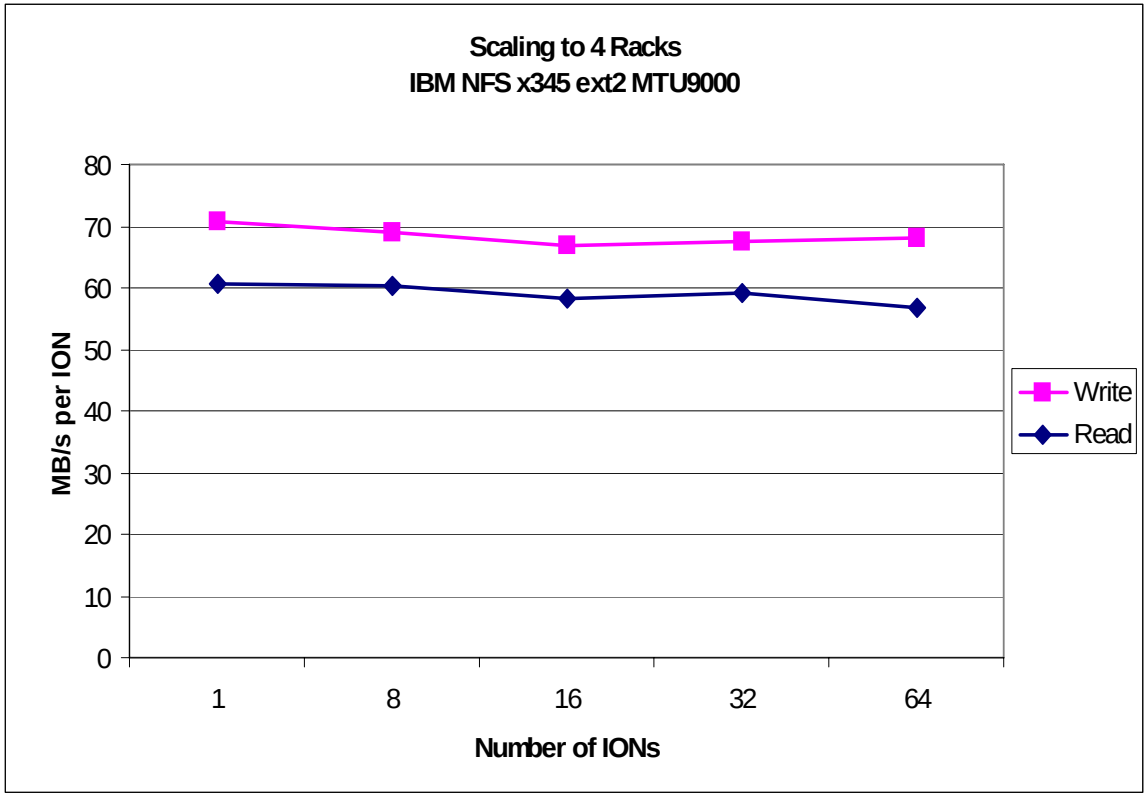
NFS Performance – Single IONode

- Single IONode with up to 64 compute nodes
- When reading, need approx 4 compute nodes to drive full IONode bandwidth



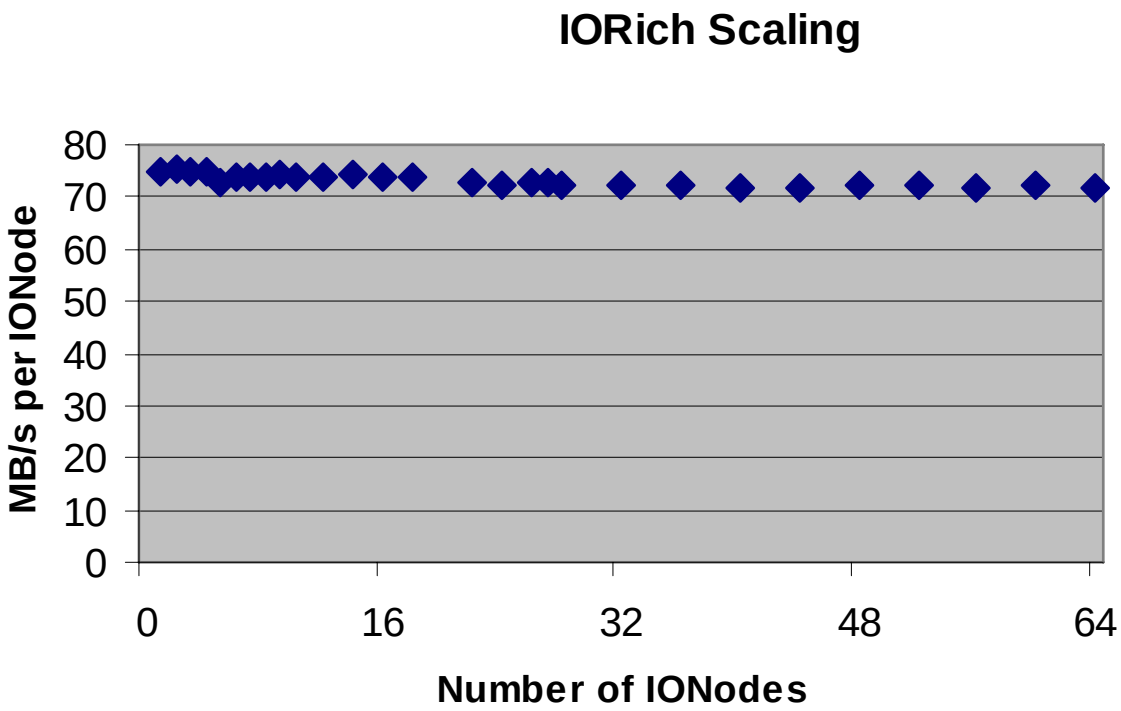
NFS Performance – Multiple IONodes

- Each IONode mounts its own server and supports 64 compute nodes
- Data points for 1 ION case, 8 IONs (midplane), 16 IONs (rack), 32 IONs (2 racks), and 64 IONs (4 racks)



NFS Performance – I/O Rich Scaling

- IBM Rochester switch configuration only allowed us to benchmark a midplane of an I/O rich rack (64 I/O nodes).
- Perfect scaling until we crossed to another switch (not shown)

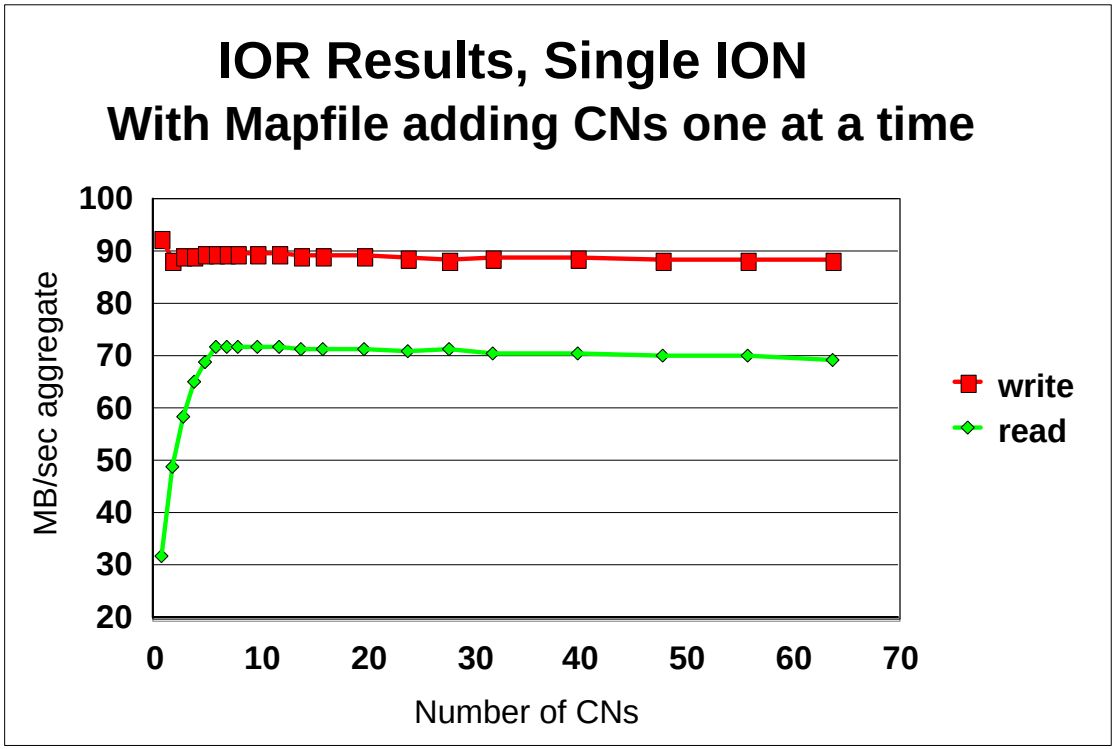


GPFS

- IBM's "General Parallel Filesystem"
- A parallel filesystem that provides a single filesystem space and can stripe data to multiple servers
- Ported to Blue Gene and delivered in release 2
- Only porting challenge was running it in an embedded environment
  - Startup is demanding on NFS where binaries and scripts reside
  - Logging and trace facilities must go to NFS
- IBM Almaden file servers were used for performance measurements
  - 32 x335 servers
  - 8 DS4300 controllers (FastT600), 1722-60U
  - 24 EXP700 Expansion drawers, 1740-1RU
  - 448 disks (each controller and drawer has 14 disks), 72 GB, 15K RPM
  - RAID-5
  - Cisco 6509 Gb Ethernet switch (8*48)

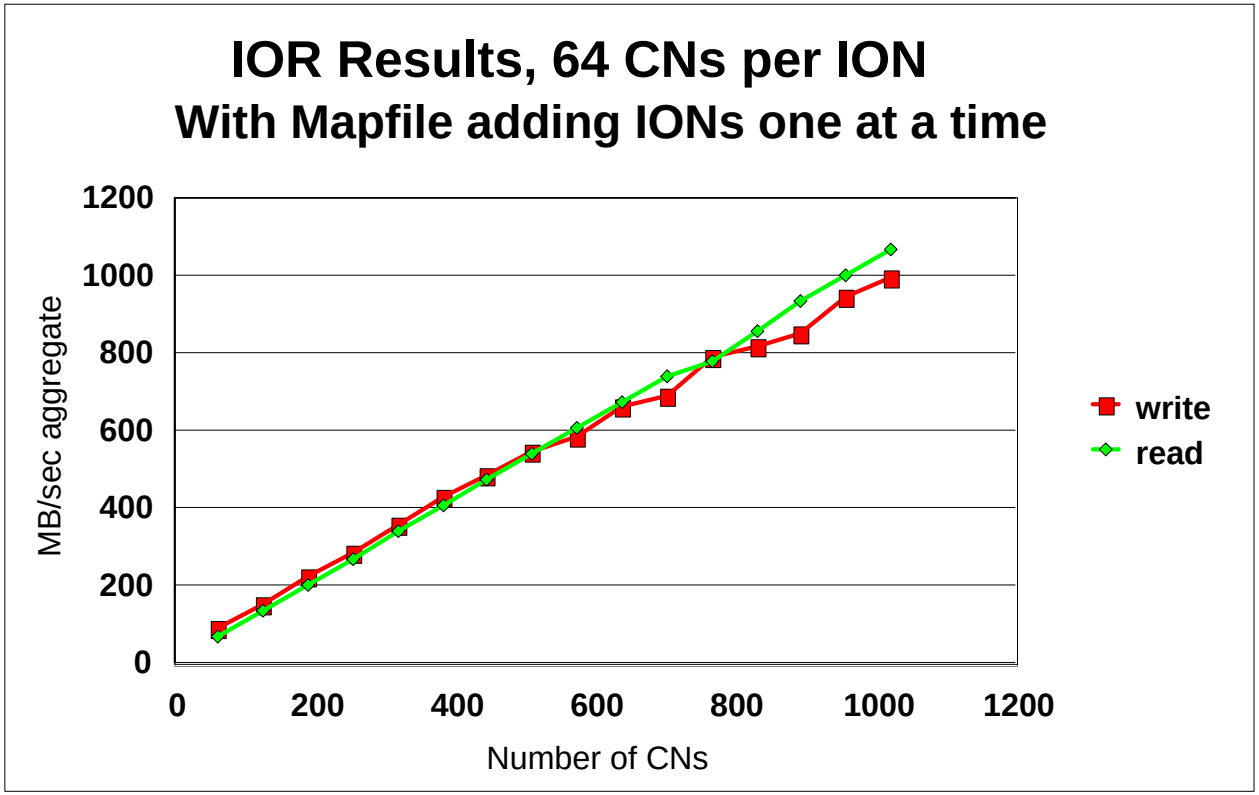
GPFS Performance – single I/O node

- Fileserver and network under test is not the bottleneck
- Read requires 6 CNs to reach peak
- Write peaks with a single CN



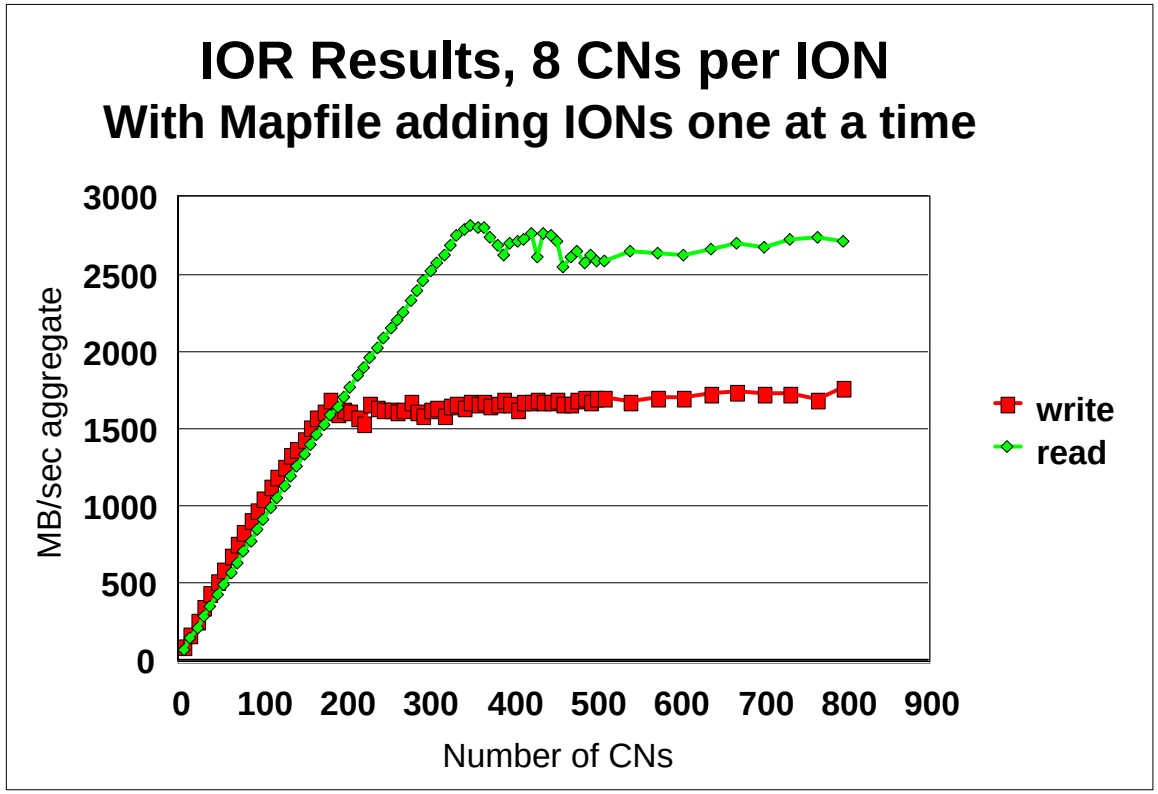
GPFS Performance – 64:1 I/O ratio

- Network and fileservers in this test were not a bottleneck
- Test results show scaling from 1..16 I/O nodes (64..1024 CNs)



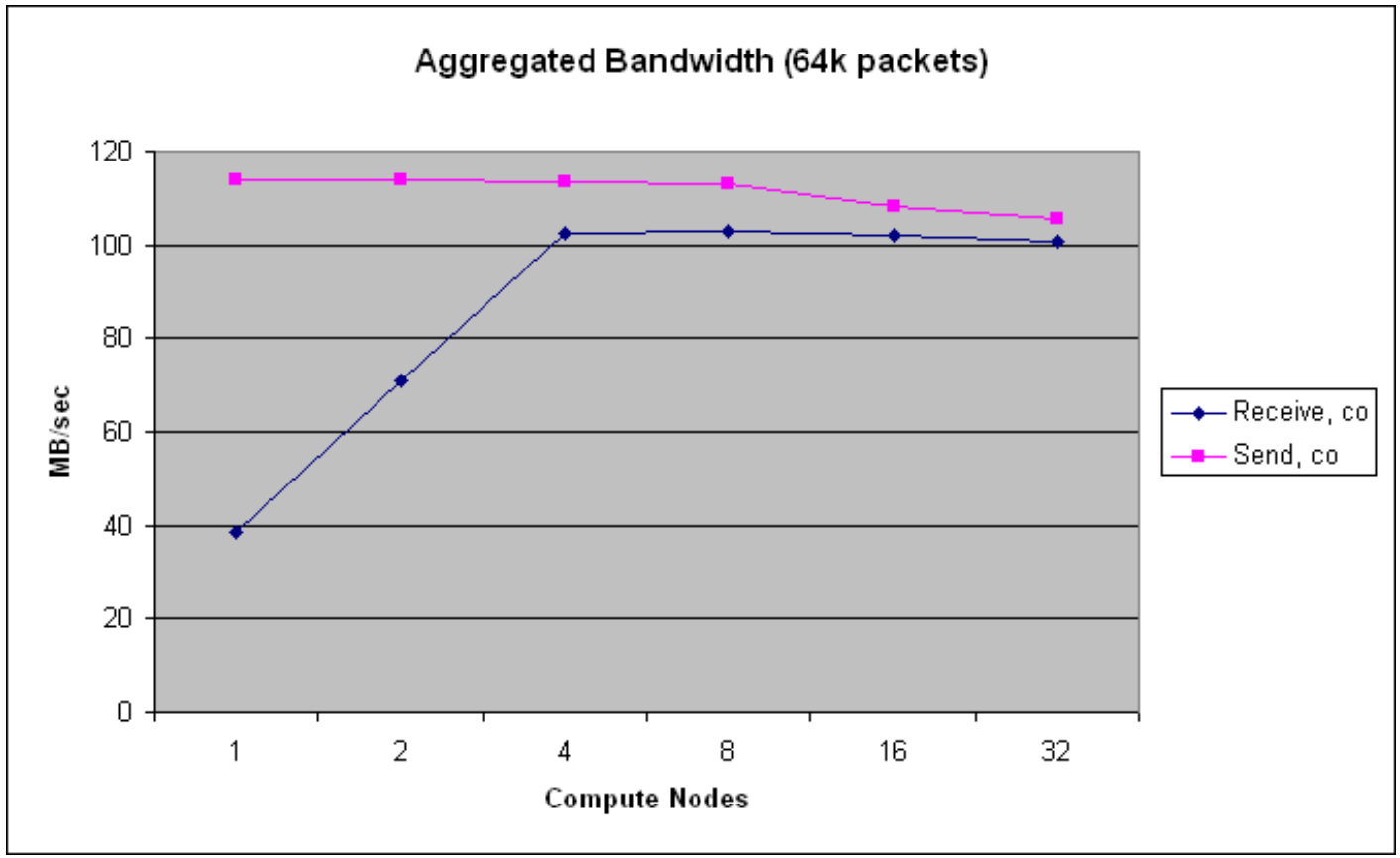
GPFS Performance – 8:1 I/O ratio

- A mapfile was used so IONodes could be brought online one at a time
- Test shows scaling from 1..128 I/O Nodes (8..1024 CNs)
- Write saturates early due to use of RAID5



Socket Performance

- Socket performance can easily saturate the gigabit link
- Greatest challenge to is require fewer compute nodes per pset



Release 2 and Release 3 enhancements

- GPFS
- Open source tree device driver
- File and socket IO performance enhancements
- CIOD callouts during job startup

GPFS on BG/L

- Read the BG/L GPFS HOWTO for GPFS setup
  - Available on the Blue Gene download site, with the BG/L GPFS RPMs
- Read the GPFS section in the /bgl/BlueLight/ppcfloor/docs/ionode.README
- For ease of administration,
  - The IONodes are in one GPFS cluster (called the bgIO cluster)
  - The NSD server nodes are in another (called the gpfsNSD cluster). This is a typical GPFS cluster.
  - Nodes in the bgIO cluster use the remote cluster capability of GPFS to mount file systems in the gpfsNSD cluster
- GPFS uses ssh and scp between nodes (requires ssh setup)
- There is a set of GPFS RPMs for the IONode
- Once installed and set up, as described in the HOWTO, just turn it on in your S10sitefs script:
  - echo "GPFS_STARTUP=1" >> /etc/sysconfig/gpfs
  - This will cause ssh and the GPFS client to start up during IONode boot
  - By default,
    - the service node is the cluster config server node
    - /bgl/gpfsvar is where the IONode-specific configs, logs, traces, etc. are located.
- CIOD_RDWR_BUFFER_SIZE should be a multiple of the GPFS block size

Open source tree device driver

- Enables easier maintenance of customer-built Linux kernels
- Tree driver provides access to hardware regs for ciod
- Tree driver implements hardware signal training procedure
- Should prove useful to ANL for ZeptoOS research

File and socket IO performance enhancements

- In the previous implementation, during a read operation data was copied from the tree hardware to a kernel buffer and then copied into the user buffer
- Data is now copied directly from the tree hardware to the user buffer
- Implemented to solve the single CN socket performance issue but also will solve the same problem for file reads

CIOD callouts during job startup

- Request from LLNL to allow file system code to register a program to be called by CIOD on job startup
- Parameters passed to program include
  - Job mode – virtual node mode vs coprocessor mode
  - Partition size
  - Job ID
- This feature can allow automatic I/O tuning per job
- Lustre will adjust for CN ratio for a significant performance improvement

Summary

- CNK to ciod message protocol is evolving with every release
- File and socket performance enhancements eliminating the need for multiple compute nodes to drive I/O to saturation
- Ciod implements more tuning parameters and job start callout
- Updated Linux 2.6.5 with additional patches
- Sockets are showing excellent performance (near wire speed)
- GPFS and NFS are showing very good performance