Blue Waters I/O Performance
CUG 2013

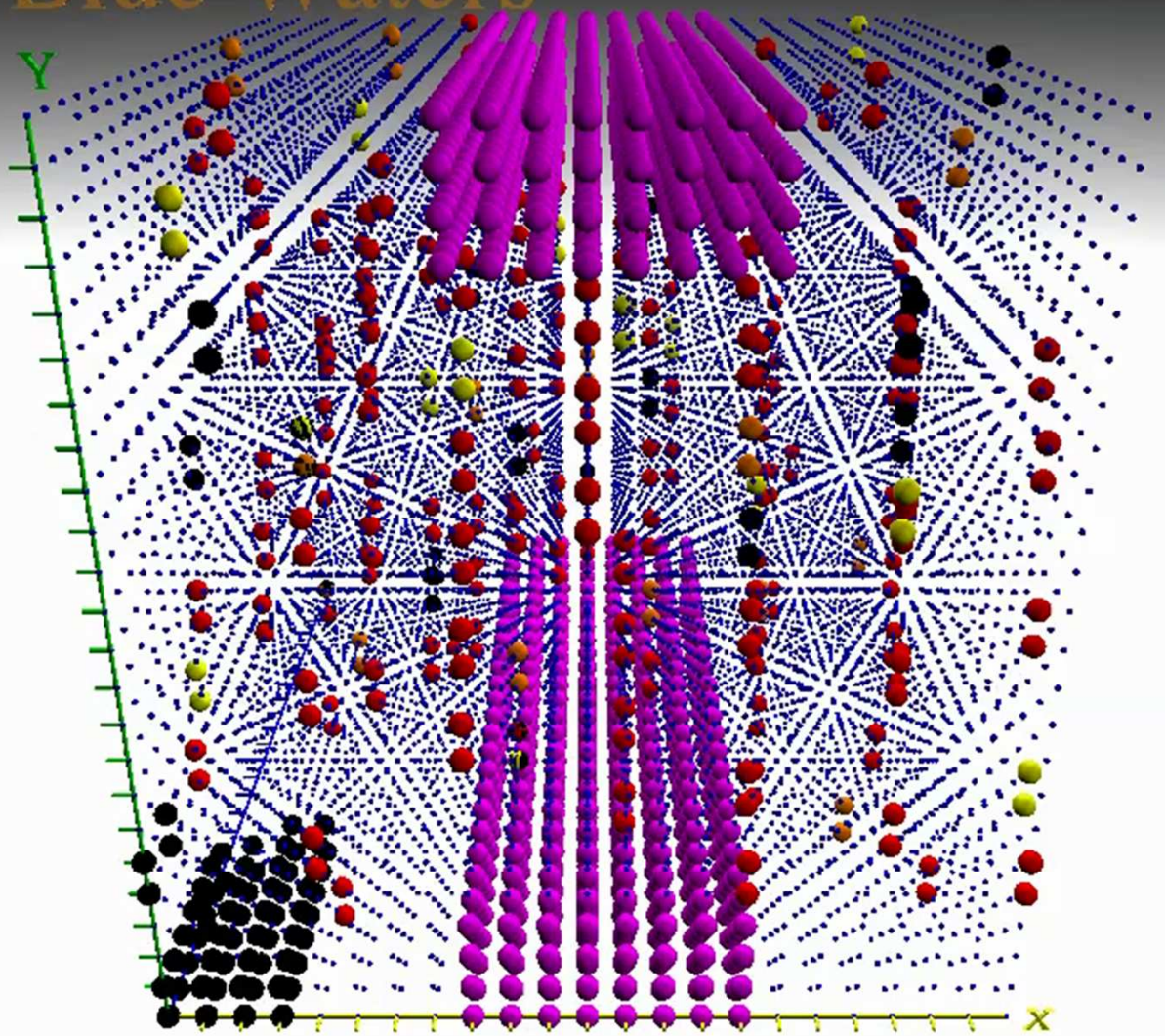
Mark Swan, Cray Inc.

Doug Petesch, Cray Inc.

What we will be talking about today

- Brief mainframe inventory
- Brief file system inventory
- LNET Fine Grained Routing (FGR) implementation
- Performance achievements
- Performance challenges
- Future explorations

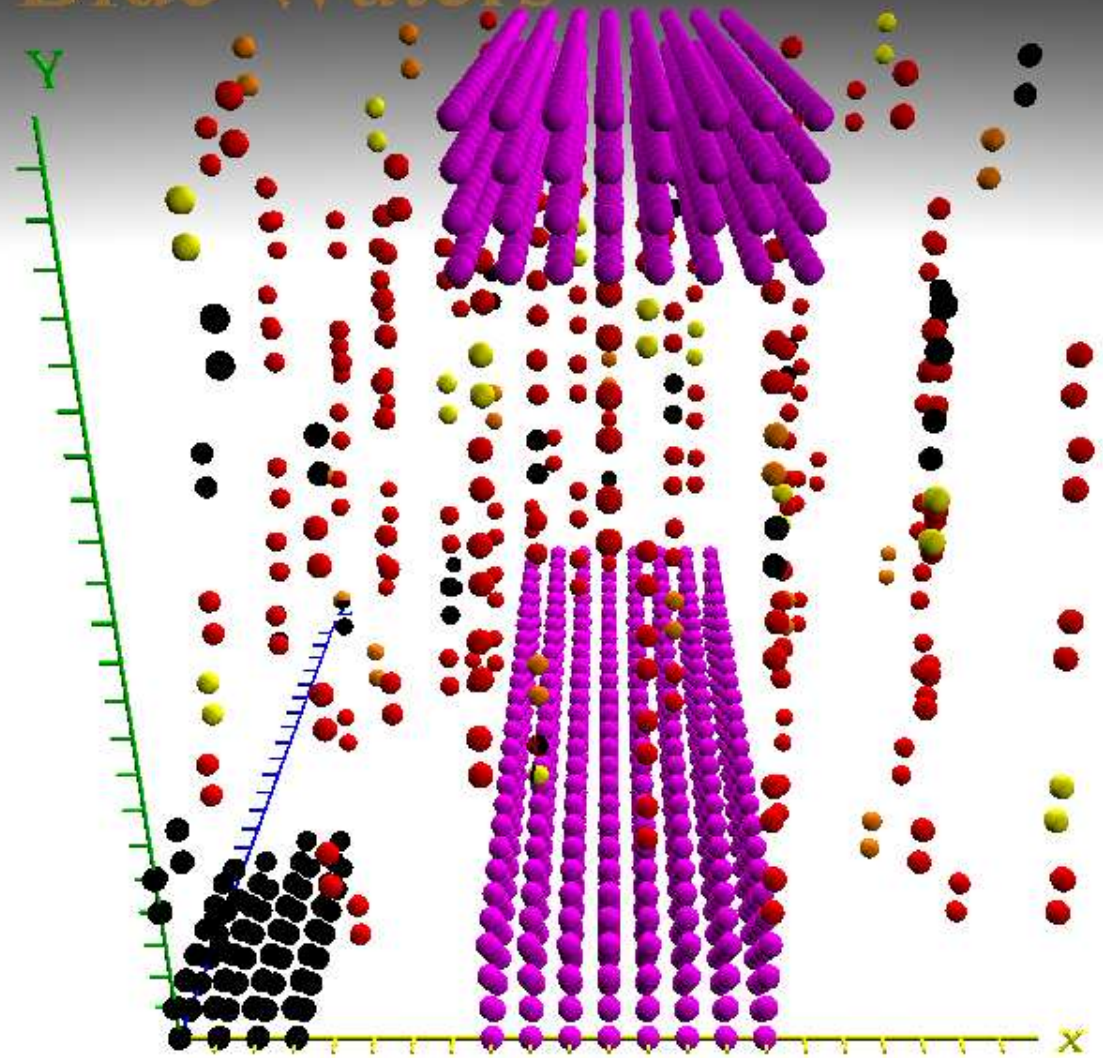
Blue Waters



'home' LNET routers
'project' LNET routers
'scratch' LNET routers

'XE6' compute
'XK7' compute
other service nodes

Blue Waters



'home' LNET routers
'project' LNET routers
'scratch' LNET routers

'XE6' compute
'XK7' compute
other service nodes

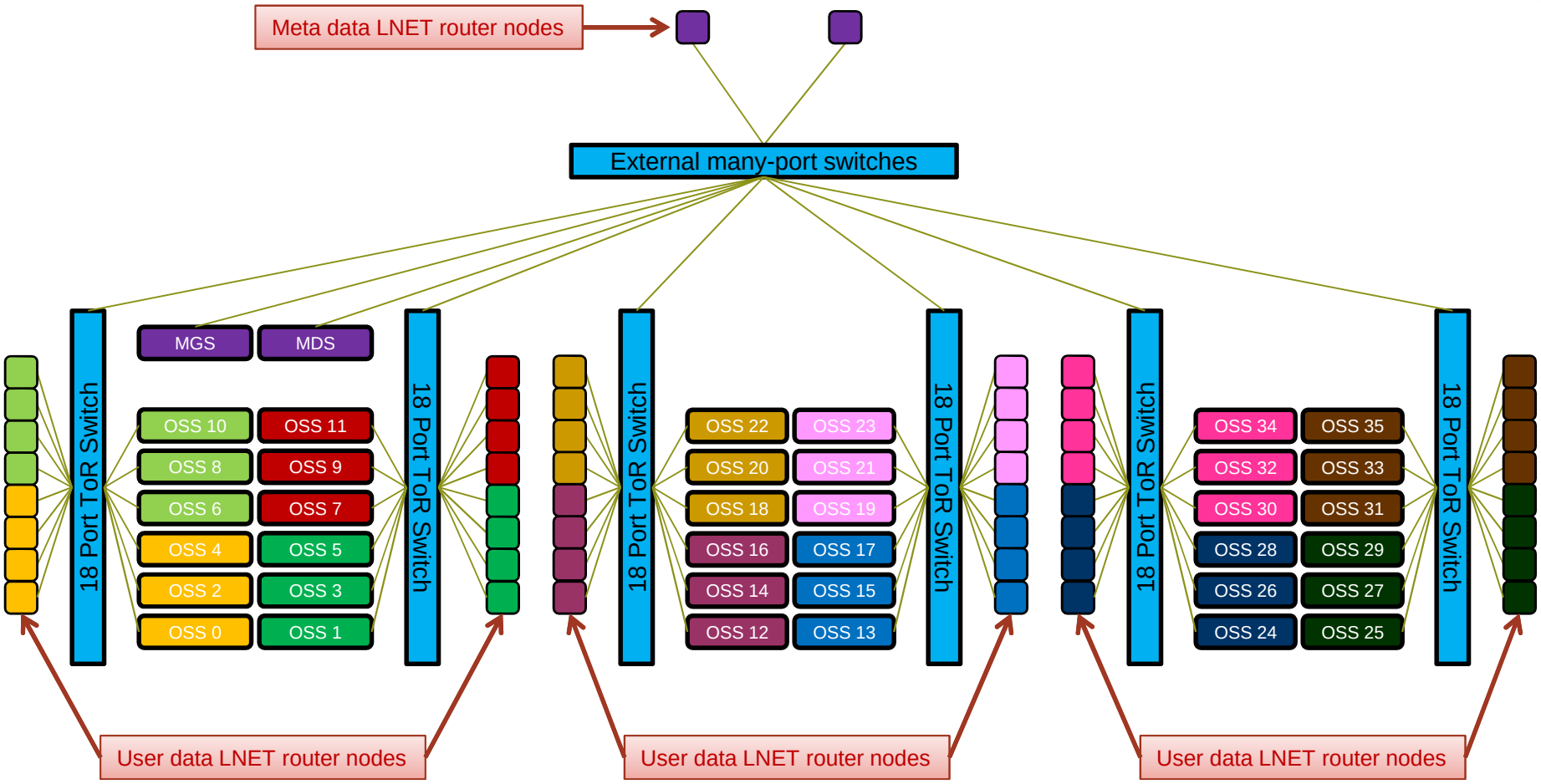
Brief mainframe inventory

- **25,712 Compute Nodes**
  - 22,640 XE
  - 3,072 XK
- **784 Service Nodes**
  - 582 LNET router nodes
    - 2 router nodes for meta-data for each file system
    - 48 OSS router nodes for “home”
    - 48 OSS router nodes for “projects”
    - 480 OSS router nodes for “scratch”
  - 202 other nodes

Brief file system inventory

- **Three Cray Sonexion 1600 file systems**
  - 18 SSU “home”
  - 18 SSU “projects”
  - 180 SSU “scratch”
- **Fine Grained Routing using a 4:3 ratio**
  - Four LNET router nodes per LNET group
  - Three OSSs per LNET group
- **All four LNET router nodes for a LNET group on same XIO blade**

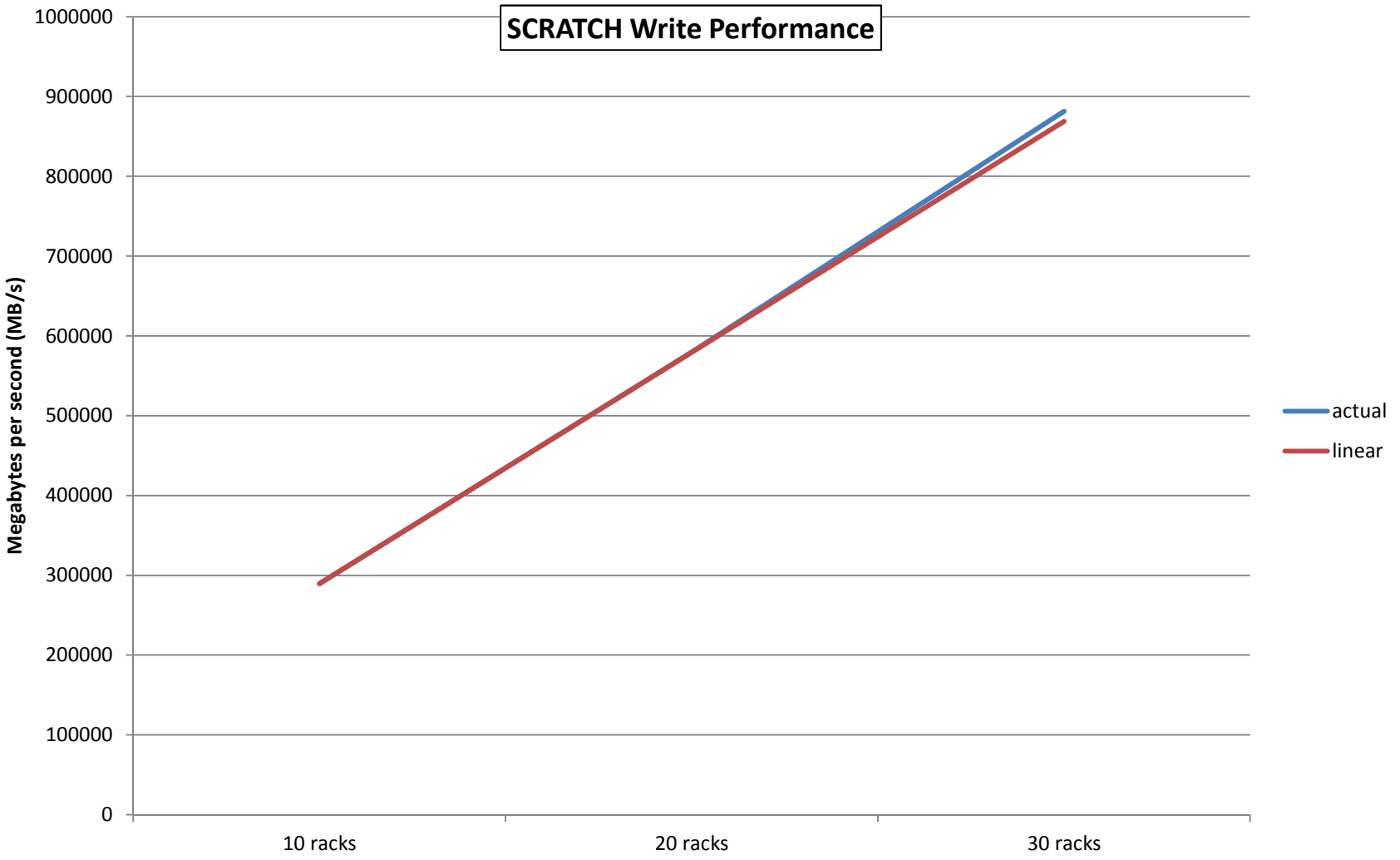
LNET Fine Grained Routing (FGR) implementation



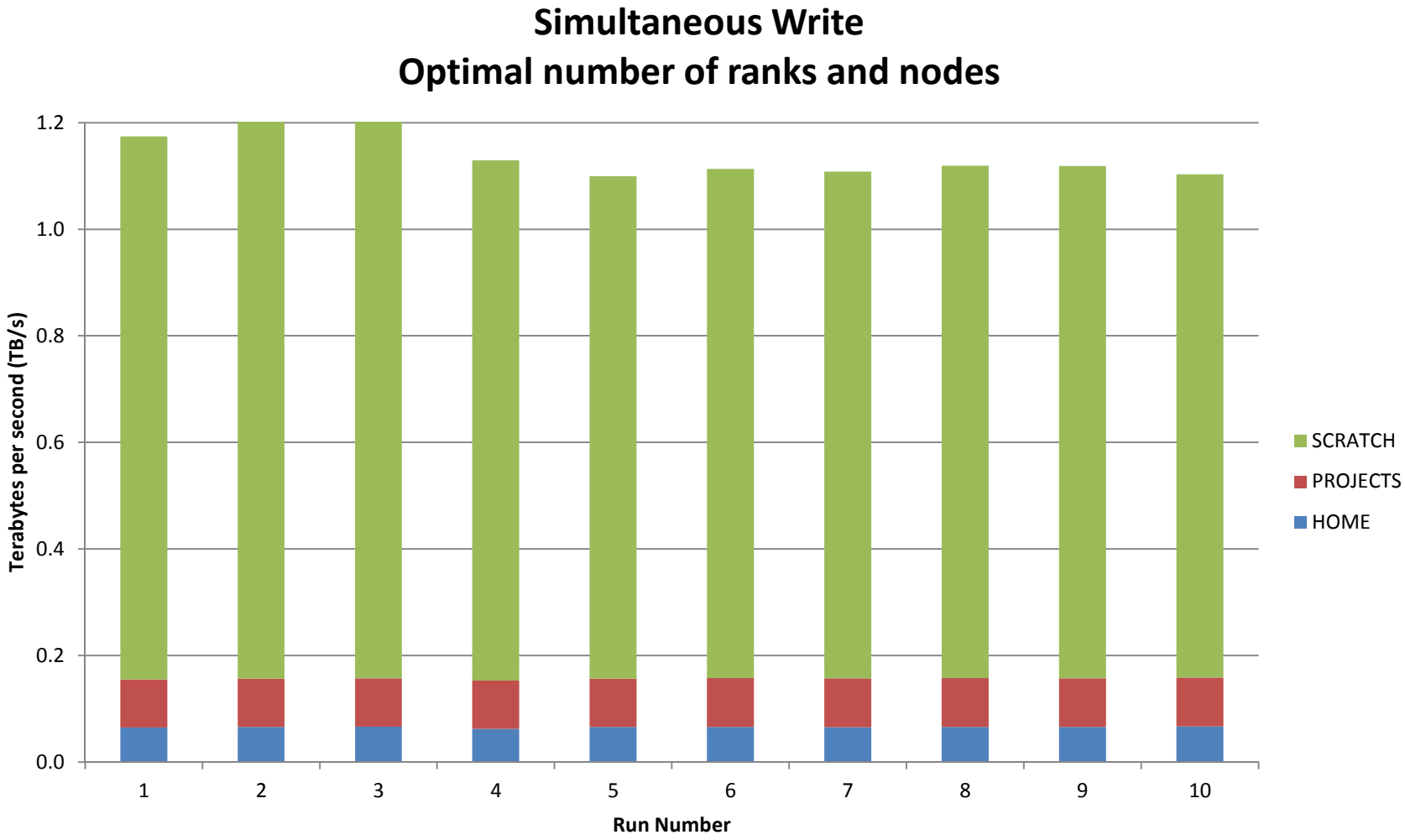
Performance Achievements

- Linear scaling
- Aggregate performance across all file systems
- Shared File MPI I/O

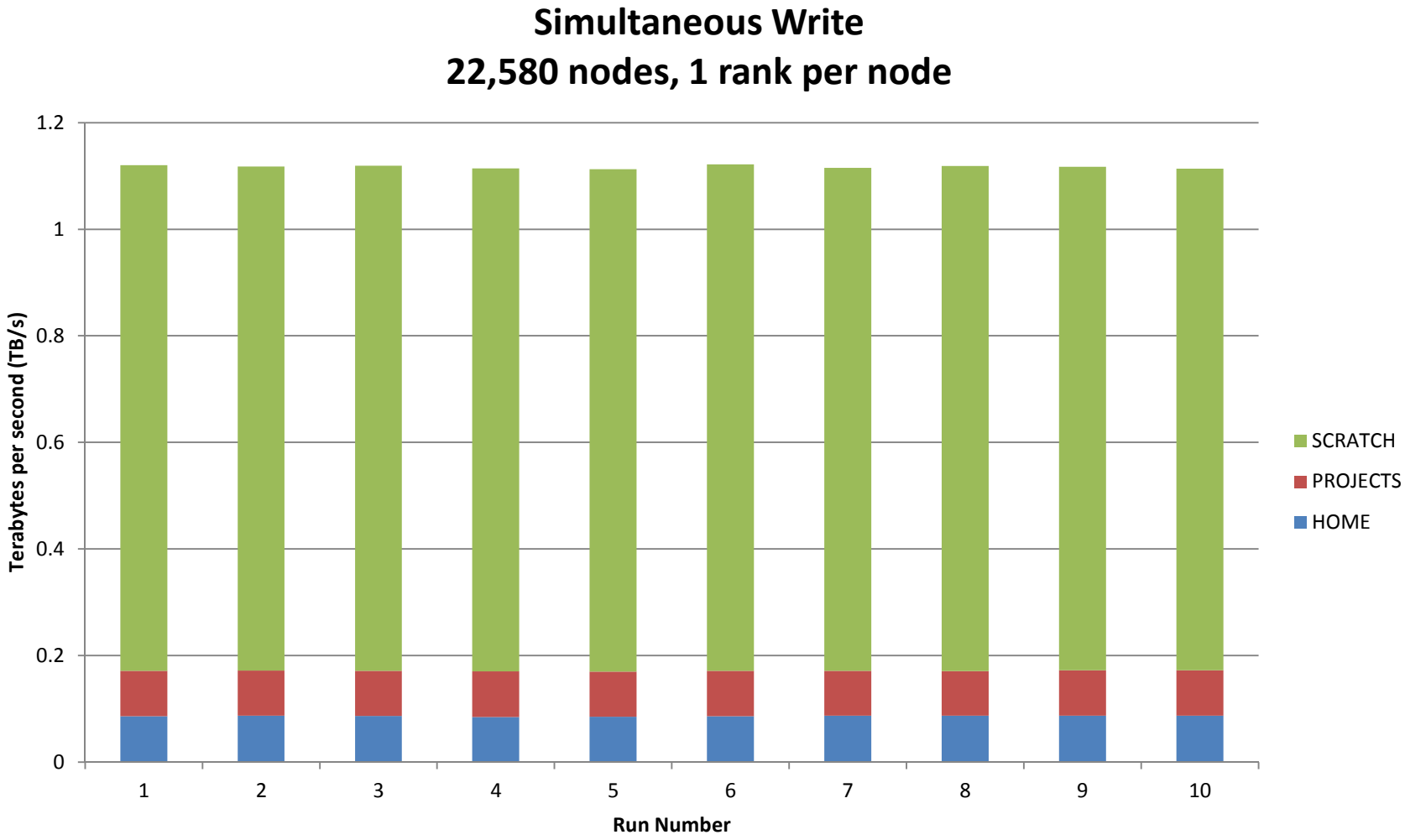
Performance achievements
Linear Write Performance



Performance Achievements
Aggregate Write Performance



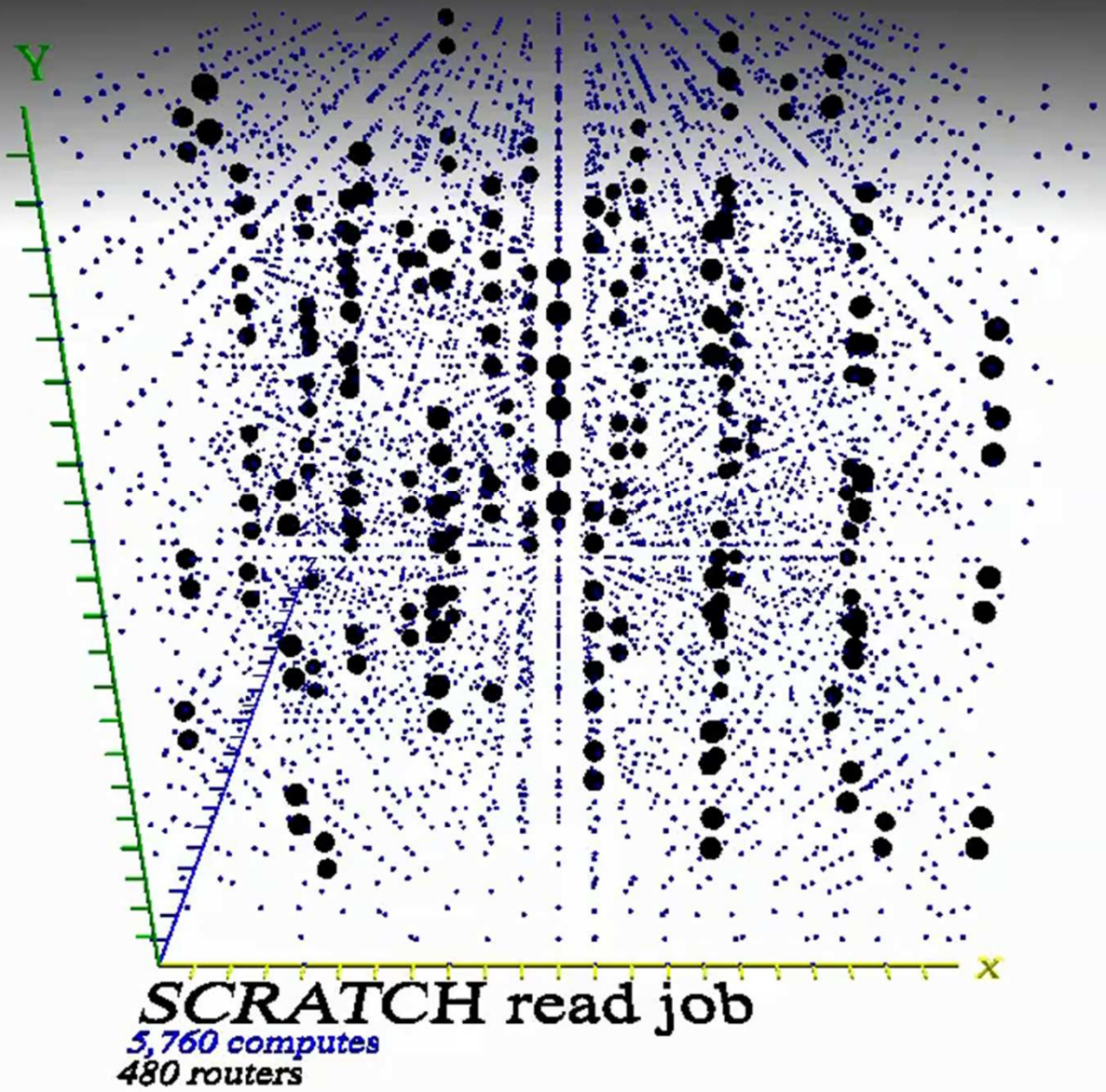
Performance Achievements
Aggregate Write Performance (cont'd)



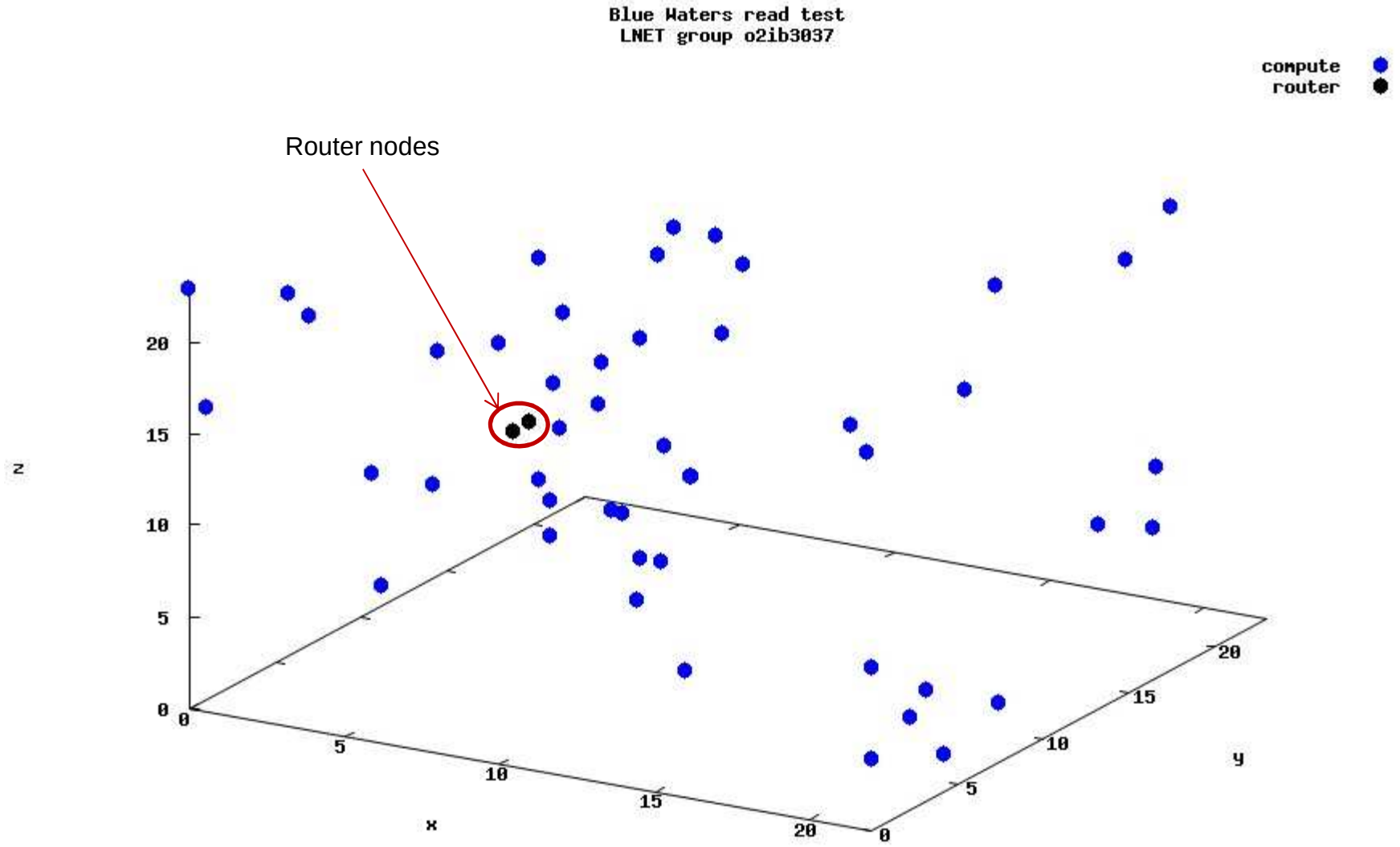
Performance Achievements
Shared File Write Performance

Performance Challenges

- **Reading striped shared files with MPI-I/O**
  - Using 2,260 ranks with stripe count of 144 is not balanced.
  - Using 2,256 ranks with stripe count of 141 is balanced.
  - Write rates improved by 12%.
  - Read rates more than doubled.
- **Fragmented file systems**
  - Cannot write large contiguous blocks.
  - Cannot read large contiguous blocks.
  - Increased head positioning for each transfer.
- **Maximizing read performance on a large Gemini**

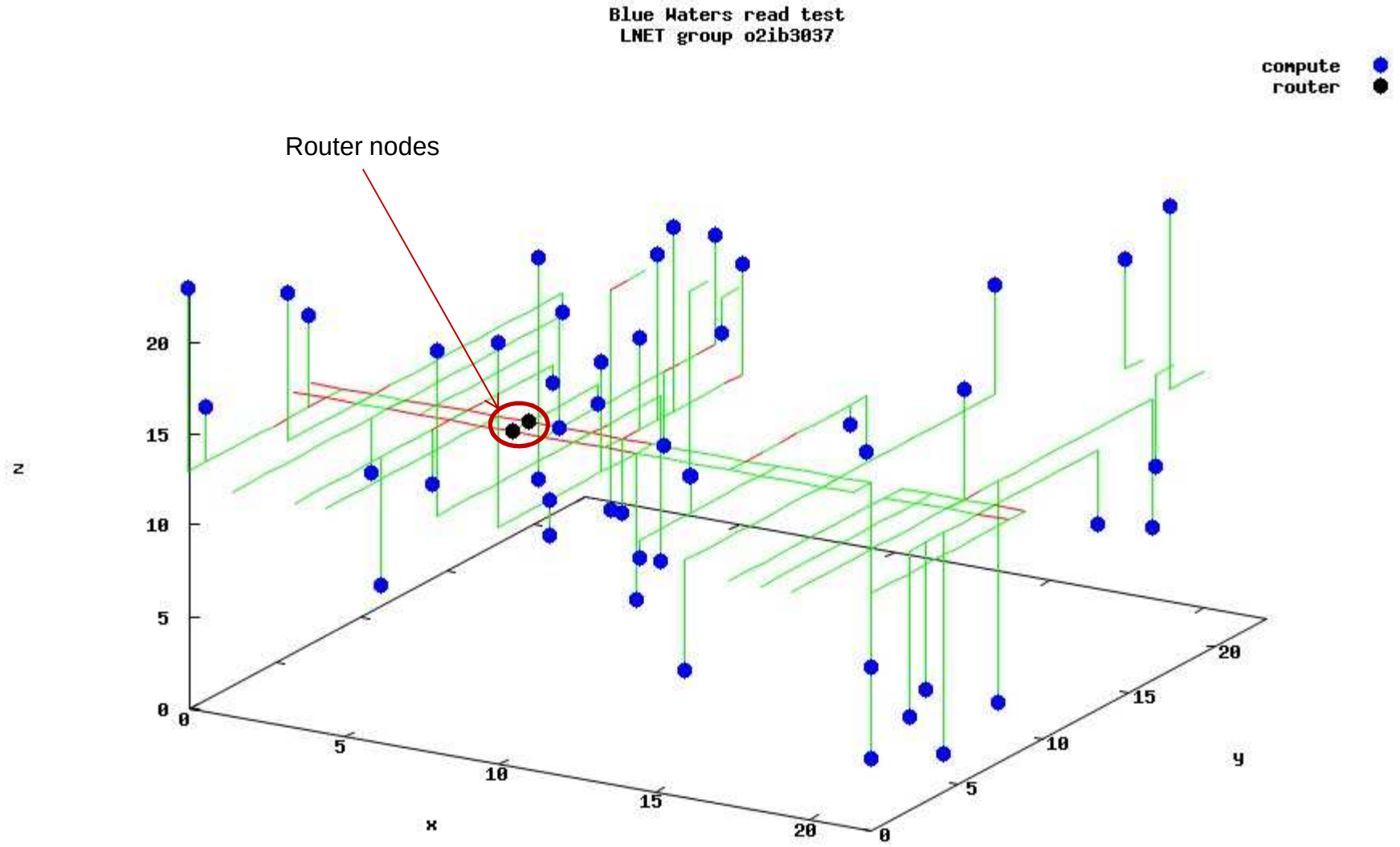

Y
X
SCRATCH read job
5,760 computes
480 routers

LNET Group o2ib3037



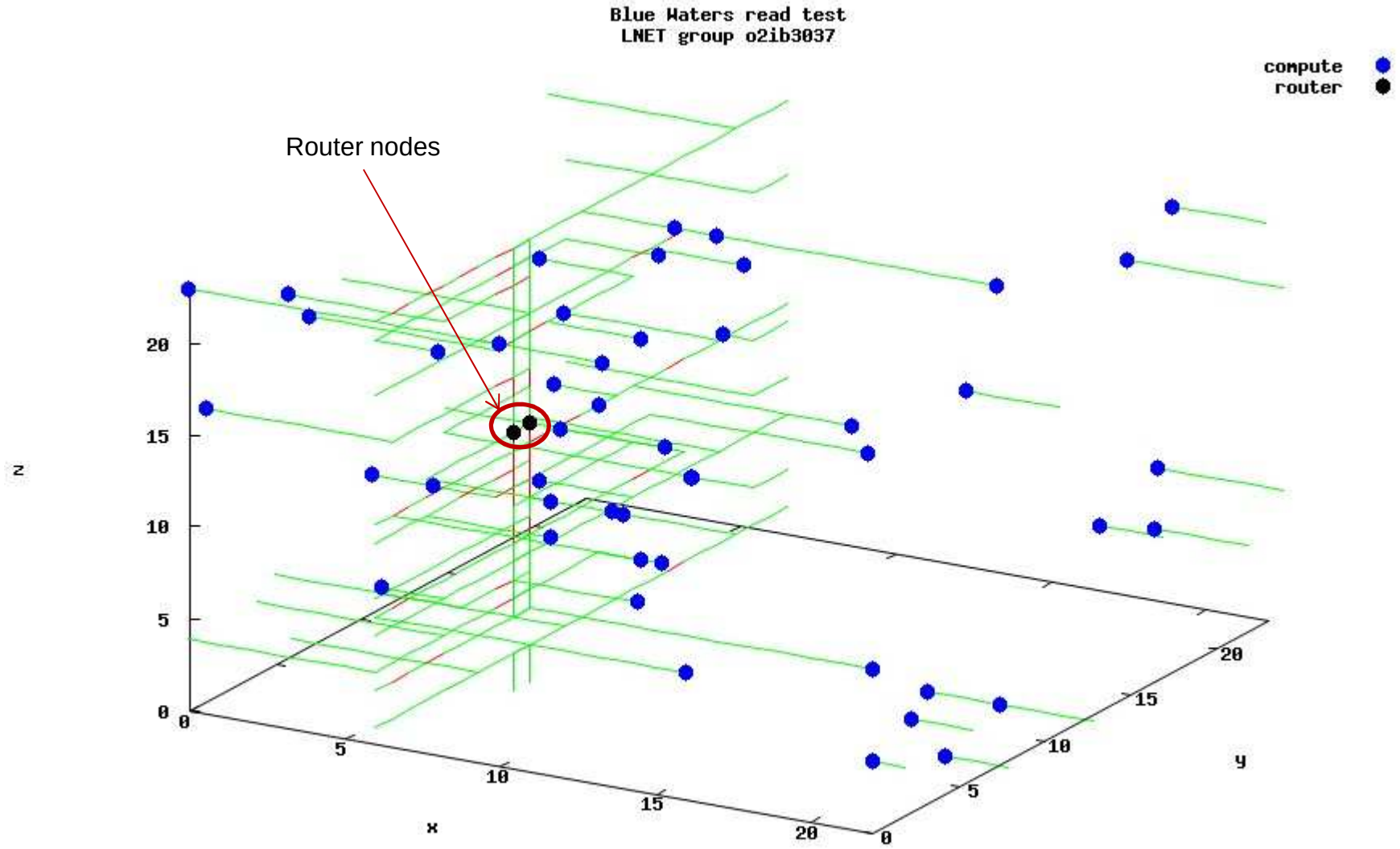
Gemini locations for the compute nodes and LNET router nodes for one LNET group.

LNET Group o2ib3037
Reading from the SCRATCH file system

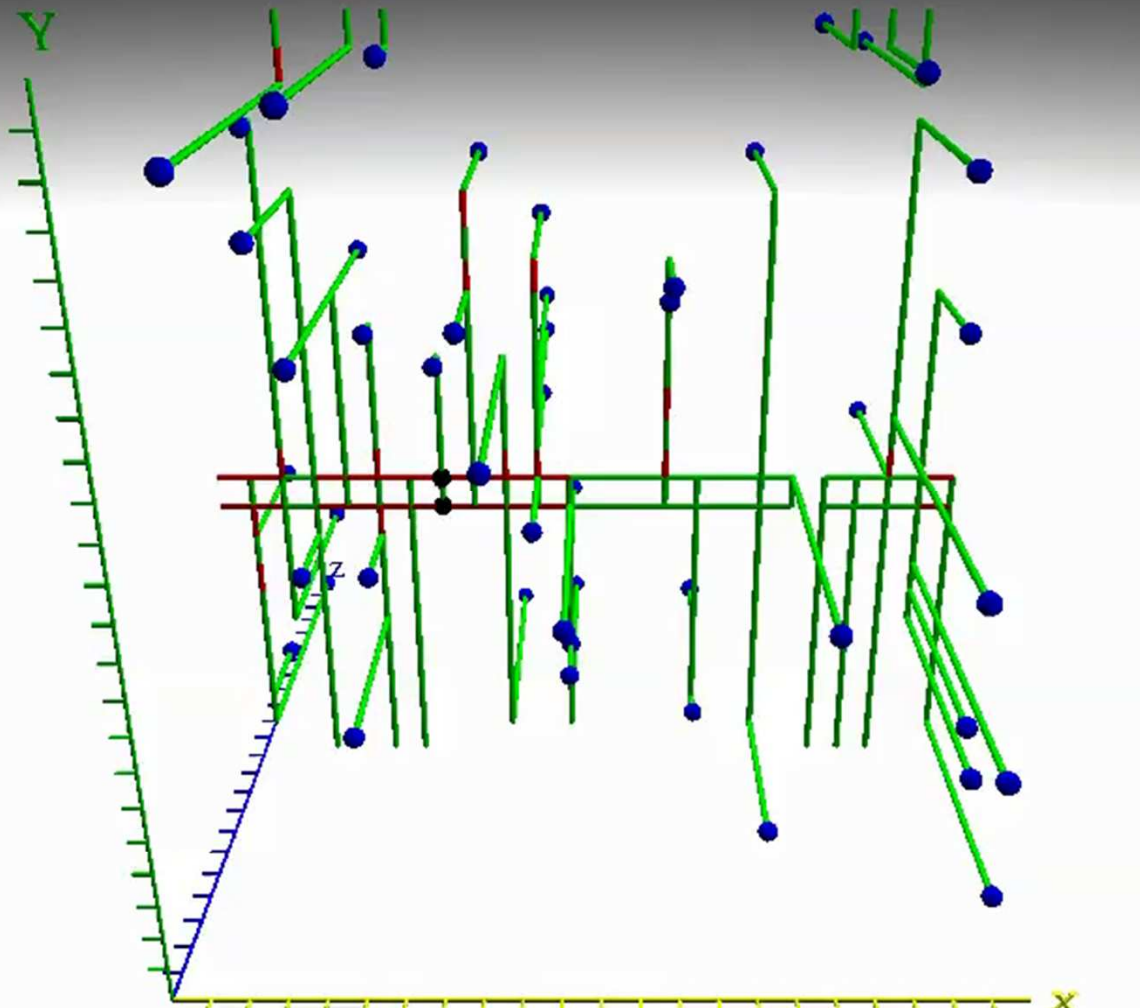


Gemini to Gemini segment loading when compute nodes are reading from the file system.

LNET Group o2ib3037
Writing to the SCRATCH file system



Gemini to Gemini segment loading when compute nodes are writing to the file system.



LNET group o2ib3037 (read)

48 computes
4 routers

Future explorations

- **Compute Node Placement for Improved I/O**
  - Job placement algorithm changes
  - Knowing which LNET router nodes service which OSTs
  - Creating halos of I/O ranks around LNET router nodes
- **OST Fragmentation and Physical Positioning**
  - Coalescing files (defragmenting)
  - Moving noncritical files to slow areas of OSTs

Thank you