Improve Large-Scale Storage System Performance via Topology-aware and Balanced Data Placement

Feiyi Wang (Ph.D)

Sarp Oral, Saurabh Gupta
Devesh Tiwari, Sudharshan S. Vazhkudai

National Center for Computational Sciences
Oak Ridge National Laboratory

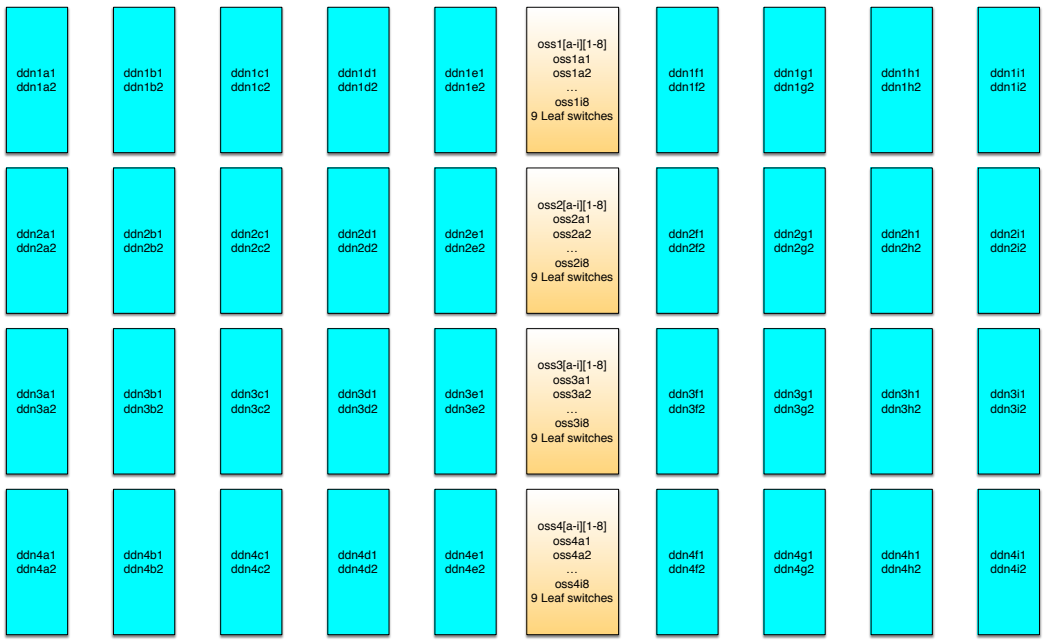
March 3, 2015

Outline

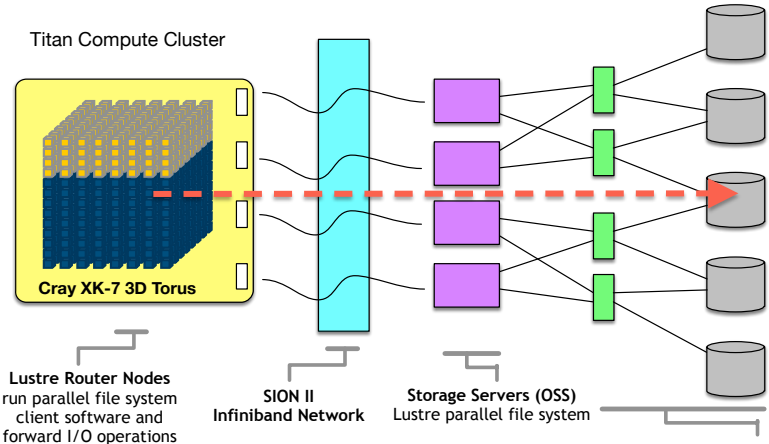
Building Storage System at Extreme Scale

- 20,160 near-line SAS drives
- 40 PB capacity (32 PB after RAID 6)
- > 1 TB/s block level performance
- 2016 Lustre OSTs
- 288 Lustre OSS
- 4 MDS and 2 MGS
- 440 Lustre LNET routers (432 for OSS, 8 for MDS)

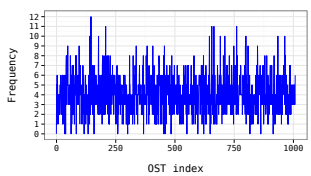
Physical Layout



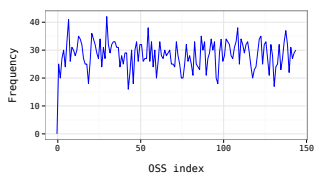
End-to-End Data Flow



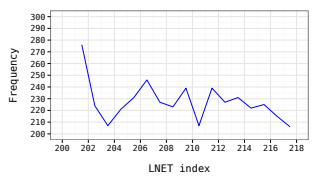
Observing the Placement Imbalance



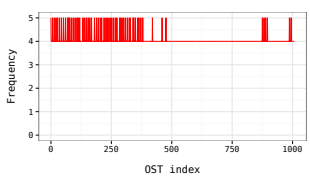
(a) Default OST



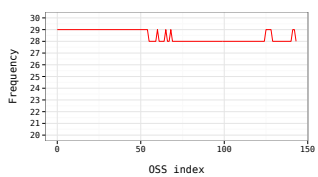
(b) Default OSS



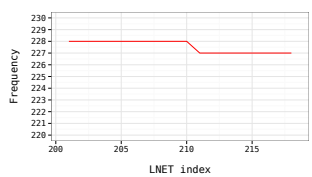
(c) Default LNET



(d) Balanced OST

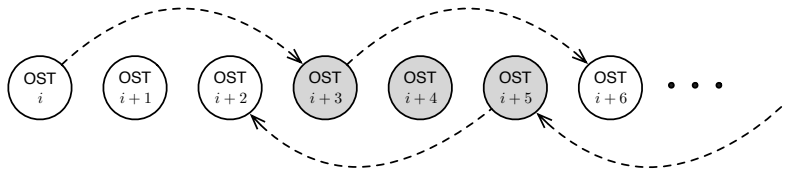


(e) Balanced OSS



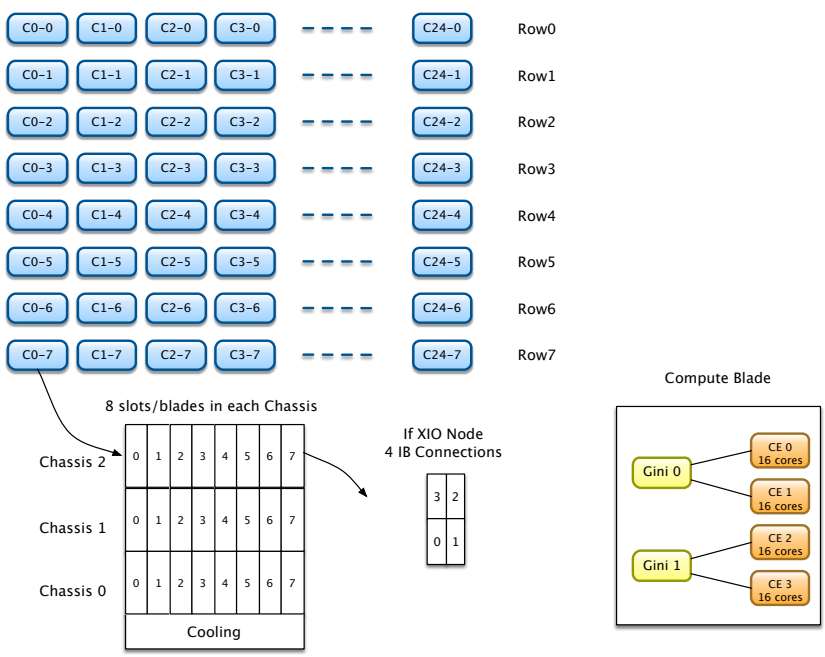
(f) Balanced LNET

Strategy Option 1: Can Parallel File System Solve It?

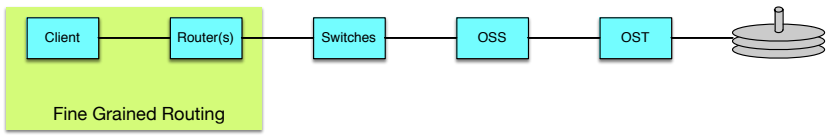


Round robin $\neq$ Load balance

Strategy Option 2: Can Infrastructure Solve It (1) ?

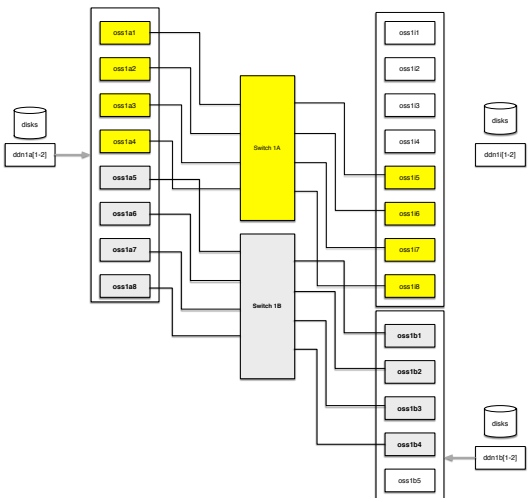


Strategy Option 2: Can Infrastructure Solve It (2)?




```
0 o2ib201:17736 o2ib202:16906 o2ib203:16930 o2ib204:1540 o2ib205:1564
o2ib206:676 o2ib207:700 o2ib208:16204 o2ib209:16180 o2ib210:17686
o2ib211:16980 o2ib212:16956 o2ib213:1626 o2ib214:1602 o2ib215:762
o2ib216:738 o2ib217:16146 o2ib218:16170 o2ib219:17737 o2ib220:16907
o2ib221:16931 o2ib222:1541 o2ib223:1565 o2ib224:677 o2ib225:701
o2ib226:16205 o2ib227:16181 o2ib228:17687 o2ib229:16981 o2ib230:16957
o2ib231:1627 o2ib232:1603 o2ib233:763 o2ib234:739 o2ib235:16147
o2ib236:16171
```

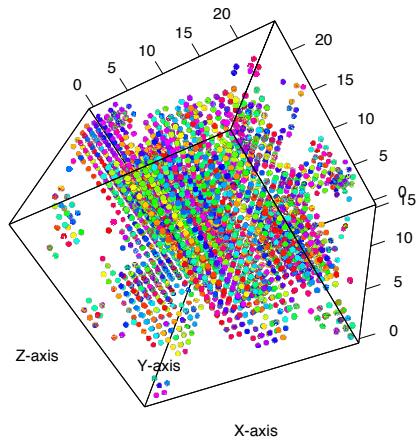

Strategy Option 3: Can Infrastructure Solve It?



OSS name	o2ib	OSTs
atlas-oss1a1	201	0, 144, 288, 432, 576, 720, 864
atlas-oss1a2	201	1, 145, 289, 433, 577, 721, 865
atlas-oss1a3	201	2, 146, 290, 434, 578, 722, 866
atlas-oss1a4	201	3, 147, 291, 435, 579, 723, 867
atlas-oss1a5	202	4, 148, 292, 436, 580, 724, 868
atlas-oss1a6	202	5, 149, 293, 437, 581, 725, 869
atlas-oss1a7	202	6, 150, 294, 438, 582, 726, 870
atlas-oss1a8	202	7, 151, 295, 439, 583, 727, 871
...		
atlas-oss2i8	210	143, 287, 431, 575, 719, 863, 1007
atlas-oss3a1	219	0, 144, 288, 432, 576, 720, 864
atlas-oss3a2	219	1, 145, 289, 433, 577, 721, 865
...		
atlas-oss4i8	228	143, 287, 431, 575, 719, 863, 1007

Job Allocation and the Need for Additional Balancing

- Full scale apps are rare.
- Cray uses ALPS (Application Level Placement Scheduler).
- Job placement is not controllable.



In-App Placement Strategy

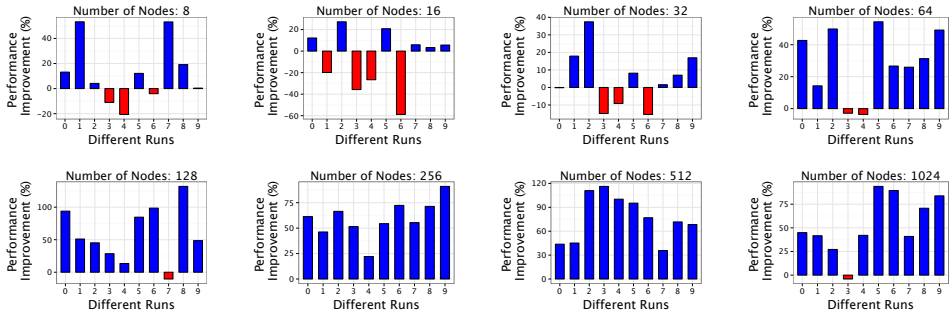
Definition

$$\begin{aligned}\text{placement cost} = w_1 \times \text{ rtr\_freq} + w_2 \times \text{ lnet\_freq} \\ + w_3 \times \text{ oss\_freq} + w_4 \times \text{ ost\_freq}\end{aligned}$$

Design Ideas

- Keep track of usage frequency for Router, LNET, OSS and OST.
- Preassign a *weight factor* for each resource involved.
- Compute placement cost for each I/O client or allocated compute nodes.
- Select the best OST and bind the I/O client with the OST.

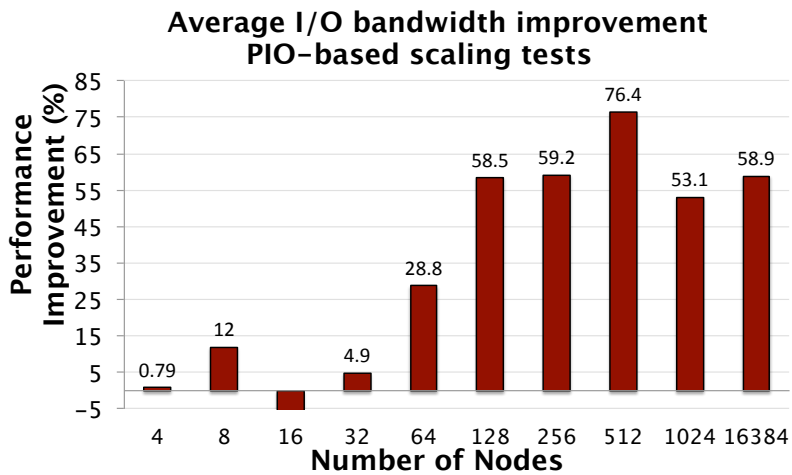
Synthetic Benchmarking Results: All Runs



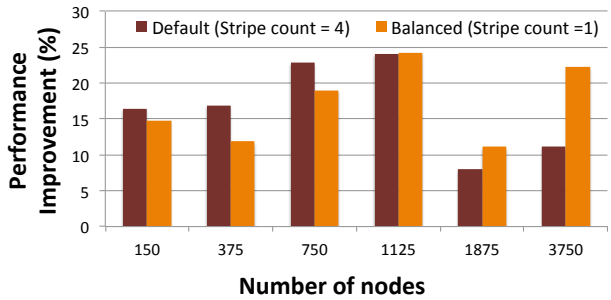
- Experiments conducted in a noisy production environment
- Multiple runs over extended period of time
- Performance improvement:

$$100 \cdot \frac{BW_{balanced}}{BW_{default}}$$

Synthetic Benchmarking Results: Average Gains



S3D Integration



- S3D is high-fidelity turbulent reacting flow solver
- S3D integration with BPIO involves 30 lines of code change
- The experiment uses all the cores present on a compute node

Conclusion and Future Work

- We analyzed a complex end-to-end data path within HPC system
- We proposed a topology-aware I/O resource allocation strategy
- We demonstrated its effectiveness through both synthetic benchmarking and a real application.

- Contention mitigation
- Systematic framework