Procuring **small** HPC storage using IO-500 benchmark

The IO-500 and the Virtual Institute of I/O BOF held in ISC18

Frankfurt, Germany

26 June 2018

George S. Markomanolis

KAUST Supercomputing Lab

Motivation - Requirements

- KAUST Supercomputing Lab has a small cluster called IBEX (~800 nodes)
  - The system is not exactly HPC oriented: long duration queues (1 month), many single node jobs, few parallel jobs.
  - Storage is really old and not a lot of space available
  - Heterogeneous system: Intel, AMD, Nvidia
- The plan is to procure 5 PB storage with HDD and 10 metadata servers with NVMe
- We require at least 12 storage nodes and 10 metadata nodes
- Planning to use BeeGFS with ZFS filesystem
- The vendors had to use part of the IO500 benchmark to evaluate their solutions
- By small procurement, we mean that cost counts…

Benchmarking

- How to download and install IO-500
- Issues:
  - System will not be available in the factory, how to execute any benchmark? They use a single storage/metadata server
  - Maybe the storage space is not enough to execute IOR for five minutes, so we demand execution for at least 1 minute.
- We provide to the vendors the following:
  - Two IOR commands for easy and hard cases to be executed on a single OSS.
  - Two mdtest commands for easy and hard cases to be used on a metadata server
  - They can use as many cores they want (on single node)
  - They have to provide us back any I/O hints (if applicable)
  - They have to provide us the results from the commands, **we are not informing** them what is an acceptable **performance**
- The winner is the one who combines best performance, best price, and complies with all the requirements

Vendors

- We received answers from four vendors, we are not going to mention their names (vendor1-4)
- Three of the vendors did report the commands and the output of the IO-500 benchmark
- Vendor1 mentioned that they disabled the Page Table Isolation (PTI) related to the Intel flaw.

Results

Vendors	# of files created per second (easy)	# of files created per second (hard)	IOR write (easy) MB/s	IOR write (hard) MB/s
Vendor 1	25751	9668	2433	1733*
Vendor 2**	13659	4867	3435	132,54
Vendor 3	34451	8071	3544	278

* Vendor1 has disabled PTI and provided enterprise hard disks NLSAS 12Gbps

** The vendor2 did not have NVMe drives available and they used SSDs.

Vendors	Hardware for single metadata server	Hardware for single storage node
Vendor 1	8 x Intel 3.2TB, NVMe, P4600	12 x 8 7.2K NLSAS 12Gbps 512e
Vendor 2	8 x 200 GB SAS SSD	60 x 8TB SATA 7.2K HDD
Vendor 3	8 x Intel 2TB, NVMe, P4500	60 x 8TB SATA 6Gb/s 7.2K RPM

Outcome

- Winner: Vendor3
- Why?
  - Vendor 1 too expensive
  - Vendor1, by using 12 HDD per storage node, requires more power consumption and more physical space.
  - Vendor3 provided overall quite good results with IO-500
  - Vendor3 complied with all the hardware and power demands

- **ToDo: Acceptance test when the hardware arrives**

- **No question received from the vendors related on how to install/run the IO-500 benchmark**

Questions?

georgios.markomanolis@kaust.edu.sa