BeeGFS tests on Low power SoC

Antonio Falabella - Luca Lama

INFN - CNAF (Bologna)

Second ASTERICS-OBELICS Workshop - 16/19 October 2017

October 12, 2017

Overview

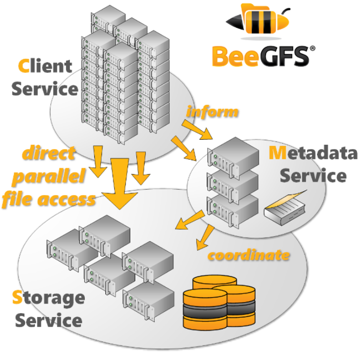
Asterics WPs involvement

- Within the working Package 3 (WP3) of OBELICS we are involved in the D-INT Asterics Deliverable D 3.9
- The idea is to benchmark the BeeGFS parallel cluster filesystem on low power SoCs (ARMv7, ARMv8 and x86).

- Why BeeGFS? Open source, scalability, easy maintenance
- Why SoC? Reduce power consumption, reduce costs, reduces sizes and weight

BeeGFS cluster filesystem

- BeeGFS is cluster parallel filesystem developed by the Fraunhofer institute (ITWM)
- It a network filesystem that allows clients to communicate with storage servers via TCP, InfiniBand, RoCE and Omni-Path



Components

- Management server
- Metadata server
- Object storage server
- Client

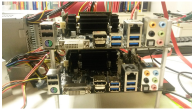
Advantages

- Open source
- Easy installation and scalability
- Doesn't require performant hardware

https://www.beegfs.io

COSA cluster at CNAF

- COSA: Computing On SoC (Systems On Chip) Architecture
- Acquire know-how on porting and benchmarking low power/low cost SoCs
- Operation of Linux systems on SoCs
- Technology tracking

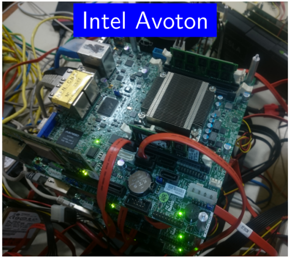


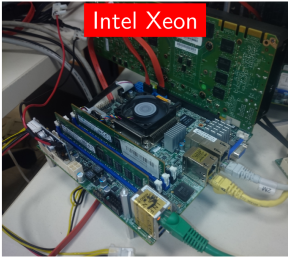
16xARMv7, 8xARMv8

4xINTEL AVOTON C-2750, 4xINTEL XEOND-1540

2xINTEL N3700, 4xINTEL N3710, 2xINTEL J4205

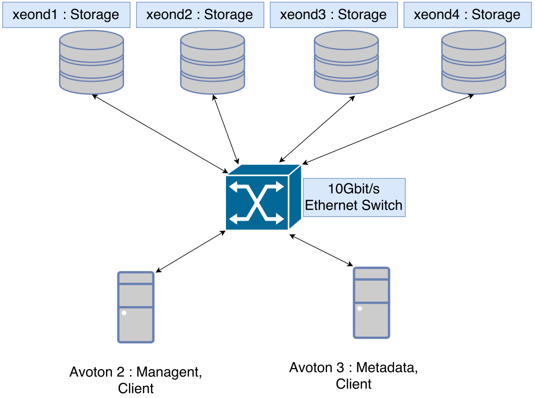
Low power SoCs





CPU	Freq	Cores	RAM	TDP	Cost
Intel Avoton C2750	2.4GHz	8	16GB	20W	100 euro
Intel Xeon D-1540	2.6GHz	8	16GB	50W	900 euro

Testbed layout



- Storage servers equipped with 8TB spinning disks
- Metadata server equipped with 500GB Samsung SSD 850 EVO

Tools

dd mainly for single process read and write tests

- `dd if=/mnt/beegfs/read_tmp_1G of=/tmp/output bs=1M count=1000`
- `dd if=/dev/zero of=/mnt/beegfs/write_tmp conv=fdatasync conv=fdatasync bs=1M count=1000`

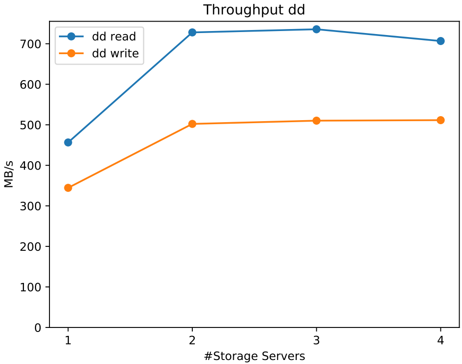
checkpoint and restart simulator with MPI

- `mpirun -np 8 allow-run-as-root python /dump_and_read.py 1000 3 10 1 1 /mnt/beegfs/ 1`

Metadata server test

- `mpirun allow-run-as-root -np 8 mdtest -d /mnt/beegfs/md/ -I 10 -z <tree depth> -b 3 -L -u -F -y -i 10`

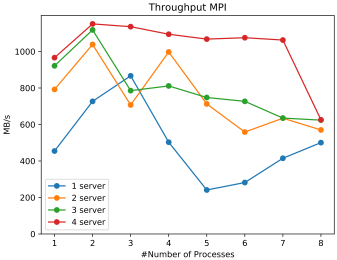
Simple dd read/write tests

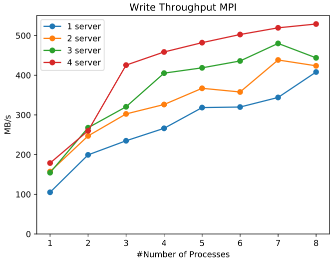


dd results

- 8 concurrent process reading and writing different files
- Throughput increases with the number of storage servers

MPI Checkpointing and restart

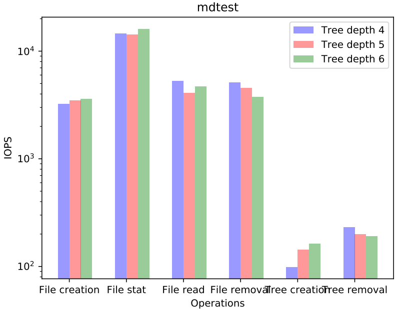




Spawns MPI jobs creating arrays of numbers of size 1G

- Read performances increase with the number of storage servers, link saturation with four
- As the number of processes increase the throughput drops due to high CPU load
- Write performances clearly show the benefits of distributed filesystem

Metadata performances

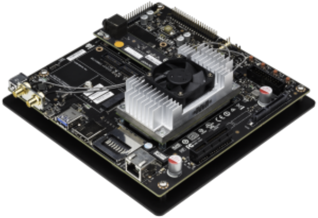


mdtest

- File stat require more IOPS has expected, CPU load not high
- Tree creation and removal not heavy, not a real removal

Further developments

- The idead is push this tests further, including ARM SoCs in the test bed such as NVIDIA Jetson T1
- 4 cores ARM A57 CPU, 1.73 GHz, 64bit
- BeeGFS software recompilation already done, to be deployed and tested
- Increase the number of clients
- Refine the measurement on caching effects

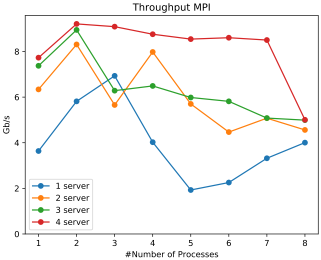


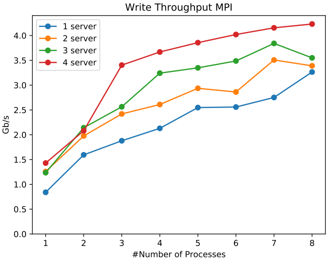
- Benchmark the filesystem on traditional servers

Summary

- At INFN CNAF within the D-INT Asterics Deliverable we are benchmarking the BeeGFS filesystem on a low power testbed
- Even if the number of machines involved in the test is not huge our tests prove that these filesystem can be deployed easily
- The resource consumption is not high, making low power SoCs a viable solution for a storage service
- We didn't spot critical issues concerning the scalability, but the number of clients should be increased
- The plan is to add clients to the testbed to stress the storage server simulating real life data access patterns

MPI Results in Gb/s





Spawns MPI jobs creating arrays of numbers of size 1G

- Read performances increase with the number of storage servers, link saturation with four
- As the number of processes increase the throughput drops due to high CPU load
- Write performances clearly show the benefits of distributed filesystem

Disks Specs

Samsung SSD 850 EVO

Performance*	Sequential Read:	Max. 540 MB/s
	Sequential Write**:	Max. 520 MB/s
	4KB Random Read (QD1):	Max. 10,000 IOPS
	4KB Random Write(QD1):	Max. 40,000 IOPS(250GB/500GB/1TB/2TB) Max. 38,000 IOPS(120GB)
	4KB Random Read(QD32):	Max. 98,000 IOPS(500GB/1TB/2TB) Max. 97,000 IOPS(250GB) Max. 94,000 IOPS(120GB)
	4KB Random Write(QD32):	Max. 90,000 IOPS(500GB/1TB/2TB) Max. 88,000 IOPS(120GB/250GB)

Seagate ST8000AS0002 8000GB SATA Hard Drive

PERFORMANCE	
Cache, Multisegmented (MB)	128
SATA Transfer Rates Supported (Gb/s)	6.0/3.0/1.5
Average Data Rate, Read/Write (MB/s)	150
Max Sustained Data Rate, OD Read (MB/s)	190