Exa-Yotta-yotta-yotta…
For Checkpoint Only

11/2008

LA-UR 08-07655

Gary Grider
Los Alamos National Laboratory

80 bytes
to
2^{80}

Questions that were posed

- Will there need to be PIS (processor in storage) in addition to PIM?
- Will there be any mechanical storage in systems of that time
- Will there be still be the three main classes for storage - scratch, persistent and archival
- What will file-systems mean in the Yotta Scale time frame
- Will storage be the weakest link of a system (reliability, latency, SW, ...)
- Will Exa-Zetta-Yotta scale data storage SW an HW be evolutionary or does it need to be revolutionary
- Will Raid-16 be enough
- Will the power profile of storage match the power profile of CPUs and memory
- How many levels will there be in the hierarchy of storage
- Will the cloud solve all the Zetta Scale storage issues
- Will tape exist in the storage hierarchy
- How much data will we access in order to read a byte? (this is page blocking, etc.)
- Will commercial needs solve the yotta scale storage issues for HPC
- Will finding the data take longer than processing the data?

Some Base Assumptions

	2008	2011	2014	2017	2020
drive size TB	1	4	16	64	256
mb/sec	50	100	200	400	800
seektime	6	4	4	3	3
power including Enc+	20	15	15	10	10
$ raw drive	200	200	200	200	200
$ drive + Enc	400	400	400	400	400
$ drive + Enc + SAN	500	500	500	500	500
SSD TB	0.05	0.5	2	8	32
MB/Sec	250	500	1000	2000	4000
$ total cost	2000	1000	1000	1000	1000
Tape TB	1	4	16	64	256
MB/Sec	150	300	600	1200	2400
$ total cost	150	150	150	150	150

Disk Based Scratch – Silly $

Big Environment	2008	2011	2014	2017	2020
pf	1	5	100	500	1000
cores (M)	0.2	1	10	50	100
memory TB	20	200	5000	10000	20000
ckpt bw TB/sec	0	1.25	20	100	400
power MW	4	8	20	40	100
File System					
drives	5000	12500	100000	250000	500000
space PB	5	50	1600	16000	128000
N dumps	250	250	320	1600	6400
N files (dumps*procs*10) (M)	500	2500	32000	800000	6400000
largest single file TB	20	200	5000	10000	20000
storage power MW	0.1	0.188	1.5	2.5	5
% of power for system	2.5	2.344	7.5	6.25	5
files for single dump (M)	0.2	1	10	50	100
checkpoint time (seconds)	80	160	250	100	50
storage $M	2.5	6.25	50	125	250

Disk Based Scratch – Realistic $

Big Environment	2008	2011	2014	2017	2020
pf	1	5	100	500	1000
cores (M)	0.2	1	10	50	100
memory TB	20	200	5000	10000	20000
ckpt bw TB/sec	0.25	1.25	8	20	48
power MW	4	8	20	40	100
File System					
drives	5000	12500	40000	50000	60000
space PB	5	50	640	3200	15360
N dumps	250	250	128	320	768
N files (dumps*procs*10) (M)	500	2500	12800	160000	768000
largest single file TB	20	200	5000	10000	20000
storage power MW	0.1	0.1875	0.6	0.5	0.6
% of power for system	2.5	2.3438	3	1.25	0.6
files for single dump	0.2	1	10	50	100
checkpoint time (seconds)	80	160	625	500	416.6667
storage $	2.5	6.25	20	25	30

SSD

To make reasonable checkpoint times – but we cant afford this either

	2008	2011	2014	2017	2020
ssd TB	0.05	0.5	2	8	32
mb/sec	250	500	1000	2000	4000
$	2000	1000	1000	1000	1000
ssds	400	2500	30000	40000	50000
space TB	20	1250	60000	320000	1600000
ckpt bw TB/sec	0.1	1.25	30	80	200
checkpoint time (seconds)	200	160	166.67	125	100
storage $ ssd only (M)	$0.80	$2.50	$30.00	$40.00	$50.00

Can SSD get this cheap $/MB/sec ?

	2008	2011	2014	2017	2020
ssd TB	0.05	0.5	2	8	32
mb/sec	250	500	1000	2000	4000
$	2000	400	50	50	50
ssds	400	2500	30000	40000	50000
space TB	20	1250	60000	320000	2E+06
ckpt bw TB/sec	0.1	1.25	30	80	200
checkpoint time (seconds)	200	160	166.667	125	100
storage $ ssd only (M)	$0.80	$1.00	$1.50	$2.00	$2.50

Tape

dumps 2 per day	2008	2011	2014	2017	2020
tapes/day	40	100	625	312.5	156.25
$/year media (M)	2.19	5.475	34.219	17.10938	8.55469

Anyone thinking $34M a year for tape?

Will we be able to store even 1 dump a day?

Exabyte 2^{60}

- We are not that far from EXA from a file system/archive system total size point of view and I don't think this is a big technical deal given capacity gains in storage devices
- Yotta for file system/archive size is worthy of thought (especially given the software issues)
- We are long way away from this for a single file
- Number of files in the trillions is daunting
  - Need multiple metadata servers to be prevalent
- Number of files/dump in the many millions is duanting
  - Need to split that up somehow or not do N to N
- How are we going to pay for all this?
- Given the inability to archive this much without a storage breakthrough, we will have to have file systems be extremely reliable, a file system outage could mean weeks of re-compute
- Will we have disk, tape
  - Solid state will overtake disk for random I/O in the coming couple of years perhaps $/seek
  - Will solid state overtake disk for BW $/TB/Sec?
  - Will solid state overtake tape for capacity / $?

What do we worry about first?

- Reliability
- Getting the right answer (SDC)
- Find ways to survive until random I/O can be put on solid state or something else
- Find ways to survive until BW can become immensely cheaper
- Find ways to survive until capacity can become immensely cheaper

- Will cloud and commercial and PIS and other buzzwords save us
  - Allies at best perhaps
  - How many Linux people will find pitty on someone that needs 2^{80}
- The sheer cost may force other non storage based solutions at least in the checkpoint regime, along with other huge issues like getting the right answer/mtti on huge machines

In Summation

Retire

Yotta 2^{80}

PS

What happened to Poor Zeta?