Data/Storage Trends

08/2018

Gary Grider

LANL

LA-UR-18-28653

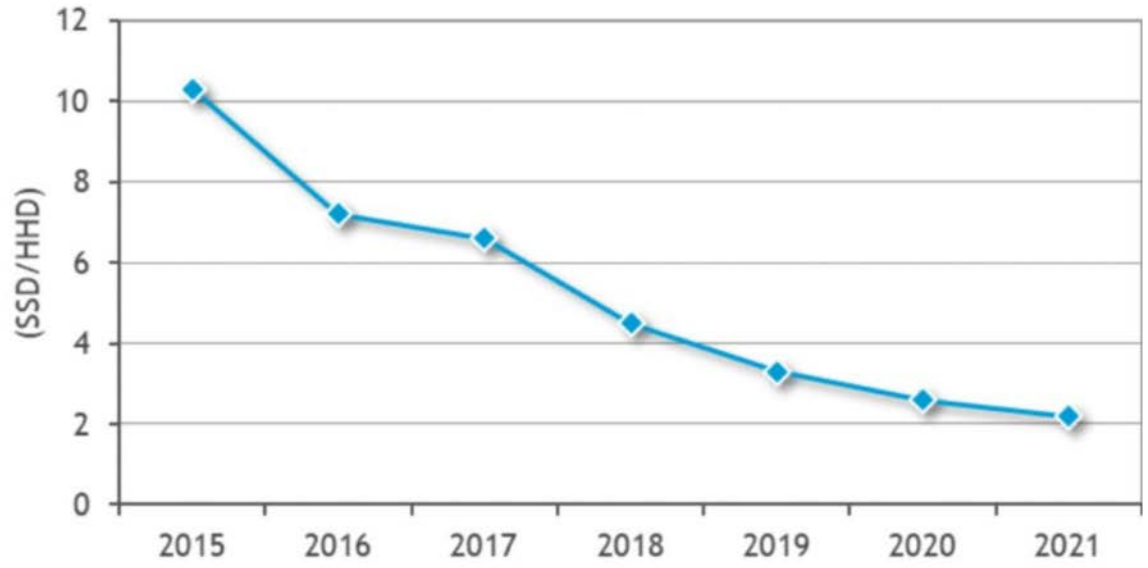
Much of this data came from a nice talk at Flash Memory Summit by

Jim Handy
(408) 356-2549
Jim.Handy(at)Objective-Analysis.com

Maybe optimistic but only a 2X premium for Flash vs Disk Capacity, do we have another "tier change happening”?

Flash Memory Summit

SSD vs HDD pricing (per gb ratio)



Source: Hyperion research
https://www.storagenewsletter.com/2018/08/07/flash-storage-trends-and-impacts/

Planar vs. 3D NAND Mfg. Cost

	16nm Planar	3D-32
Terabytes/Wafer	5.6	17.2
Wafer Cost	$1,200	$2,000
Cost/GB	$0.21	$0.12

OBJECTIVE ANALYSIS – www.OBJECTIVE-ANALYSIS.com

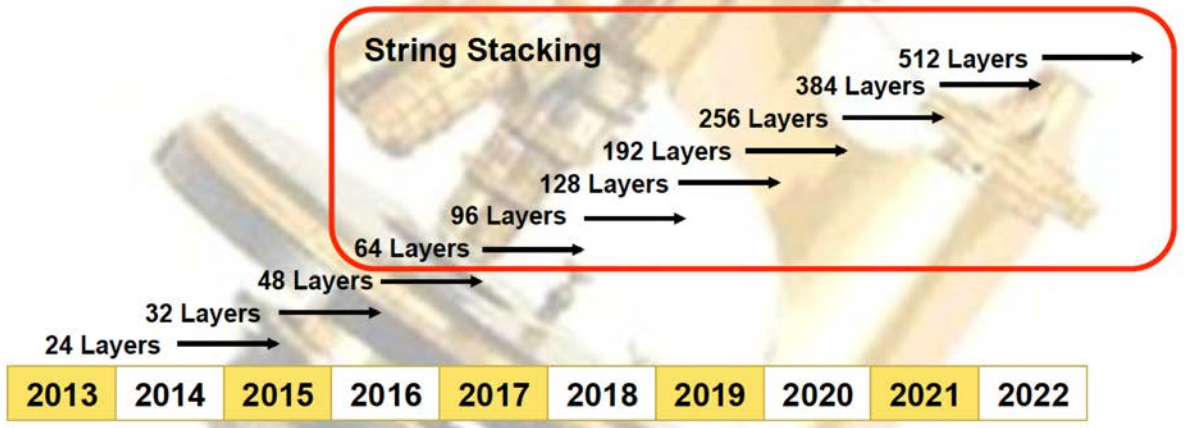
Healthy growth in market and lots of technology headroom is keeping many players engaged

Key DRAM & NAND Makers

Company	DRAM	NAND	Comments
Samsung	46%	33%	Focus: large customers & internal SSDs
SK hynix	26%	11%	Finally shipping 3D NAND in volume
Toshiba	--	19%	Spun off and ready to grow
WDC/SanDisk	--	18%	Rarely supplies chips
Micron	21%	12%	Breaking ties with Intel
Intel	--	7%	Only producing for Intel SSDs

OBJECTIVE ANALYSIS – www.OBJECTIVE-ANALYSIS.com

3D NAND Roadmap



OBJECTIVE ANALYSIS – www.OBJECTIVE-ANALYSIS.com

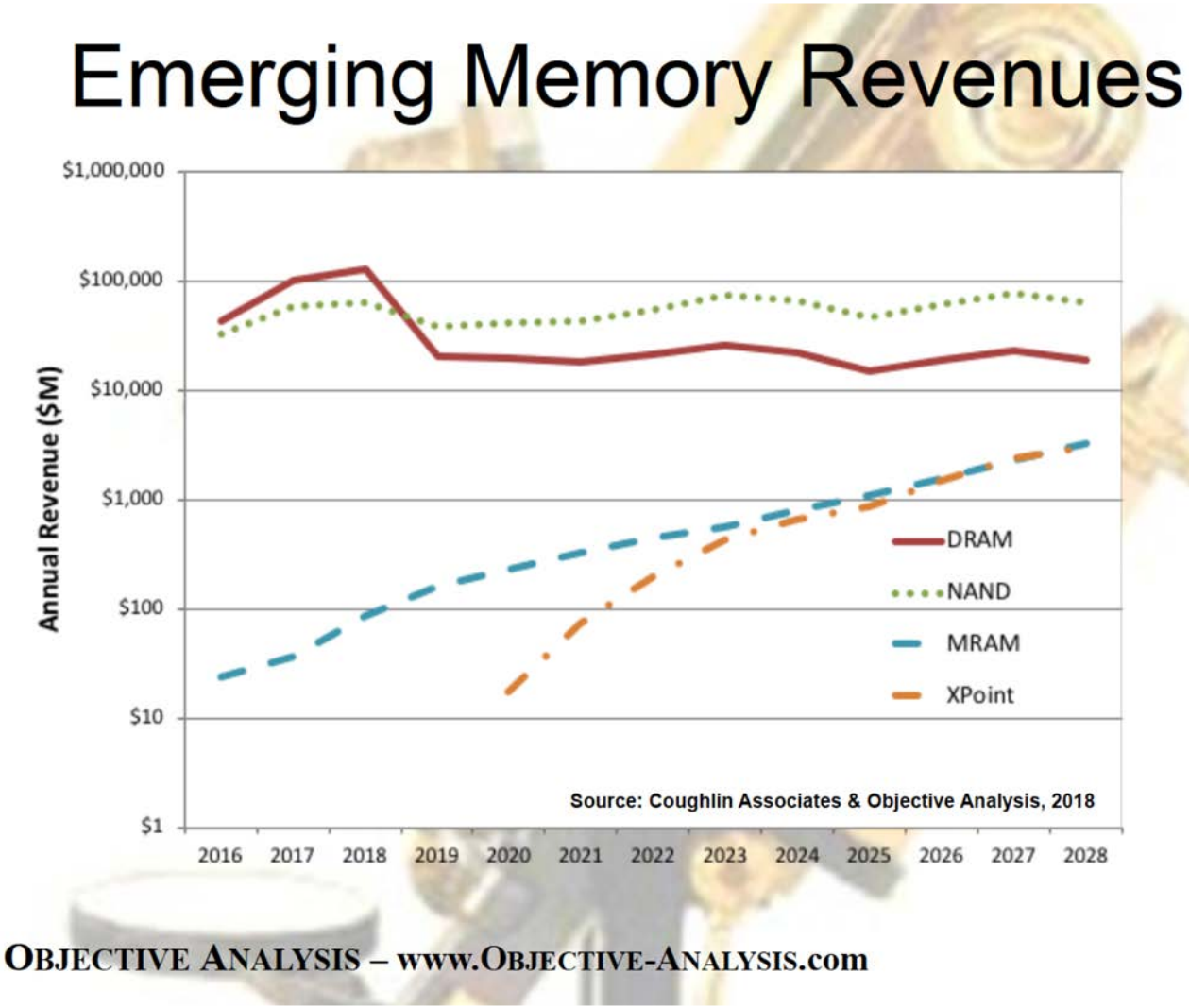
Emerging Memories Perform Better

	MRAM	ReRAM	FRAM	PCM	XPoint	NOR	NAND
Nonvolatile	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Erasable	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Programmable	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Smallest Write	Byte	Byte	Byte	Byte	Byte	Byte	Page
Smallest Read	Byte	Byte	Byte	Byte	Byte	Byte	Page
Read Speed	Fast	Fast	Fast	Fast	Fast	Fast	Slow
Write Speed	Fast	Fast	Fast	Fast	Fast	Slow	Slow
Active Power	Low	Med	Low	High	High?	Med	Med
Sleep Power	Low	Low	Low	Low	Low	Zero	Zero
Price/GB	High	High	High	High	High?	Med	V Low
Applications	Niche	TBD	Low Power	Obsolete	Main Memory	Code	Data

But when will they not be emerging anymore, Economics will guide!

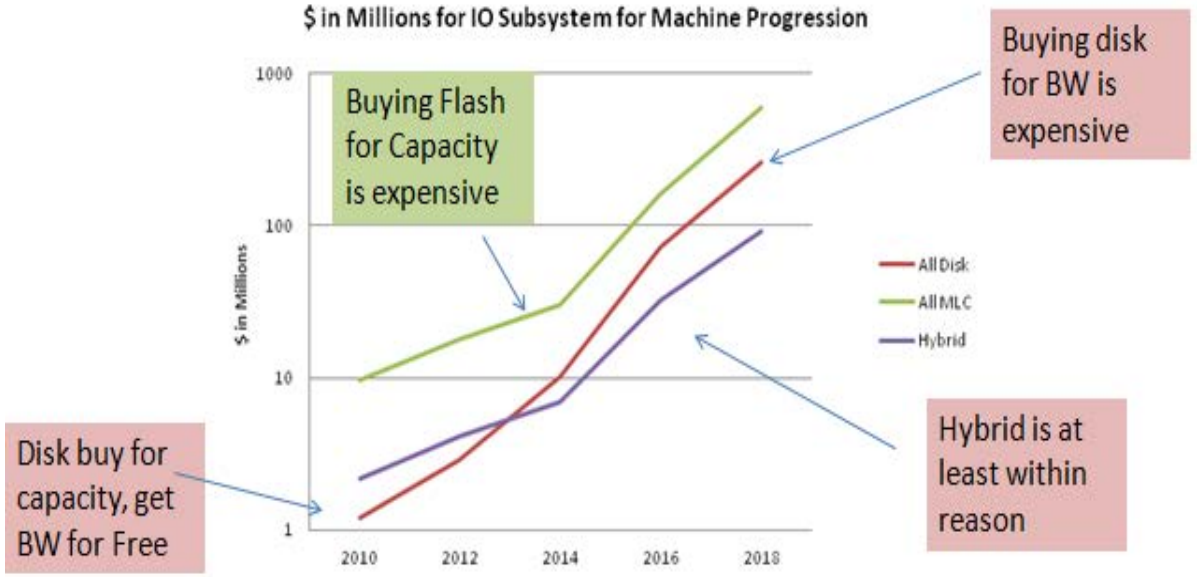
Well, maybe for some apps but not for many/most

Flash Revenues and Even margins are healthy right now but the margin part will/may change due to current shortage reactions and emerging players



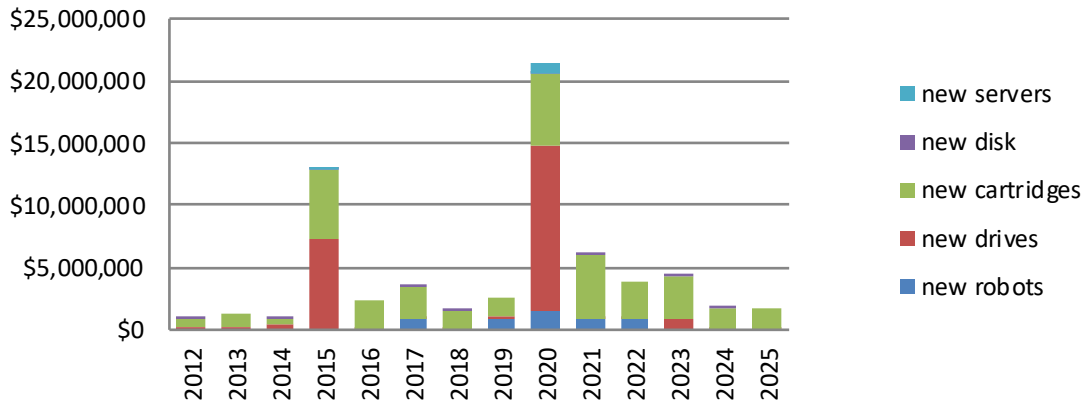
China will be a player by 2020 and this market may go the way of steel

Economics have shaped our world
Beginning of storage layer proliferation circa 2009
Disks expensive for bandwidth, tape expensive for bandwidth



Economic modeling for large burst of data from memory shows bandwidth / capacity better matched for solid state storage near the compute nodes

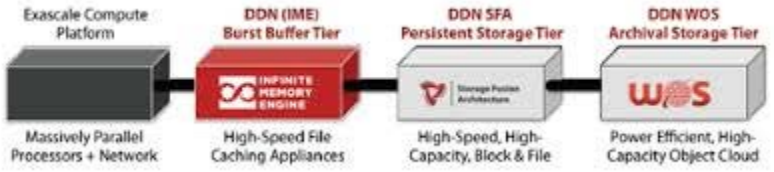
Economic modeling for archive shows bandwidth / capacity better matched for disk



Bring on the tiers – or tears
The Burst Buffer Hoopla Parade circa 2014
And Campaign Store in 2016

DDN STORAGE

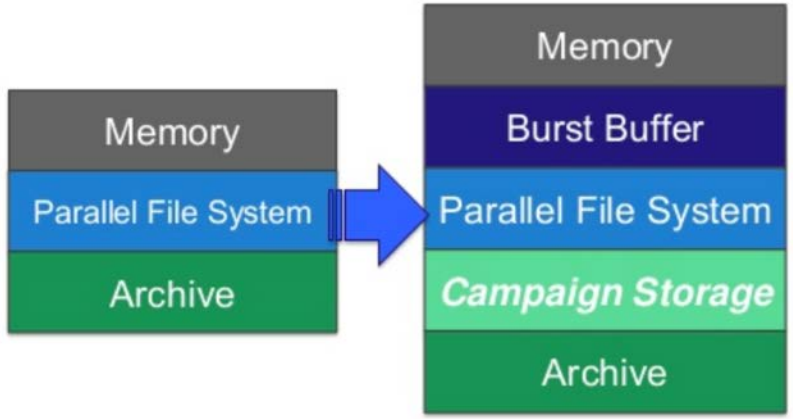
INFINITE MEMORY ENGINE™



CRAY
THE SUPERCOMPUTER COMPANY
Data Warp

Powered By
Dilithium Crystals

CS Campaign Storage

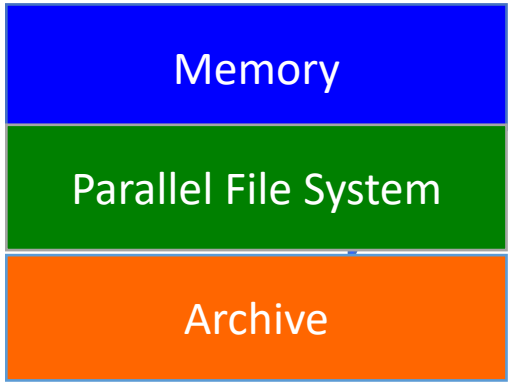


More HPC Storage products coming in the tiering space

Necessary Evil Unfortunately - Economics

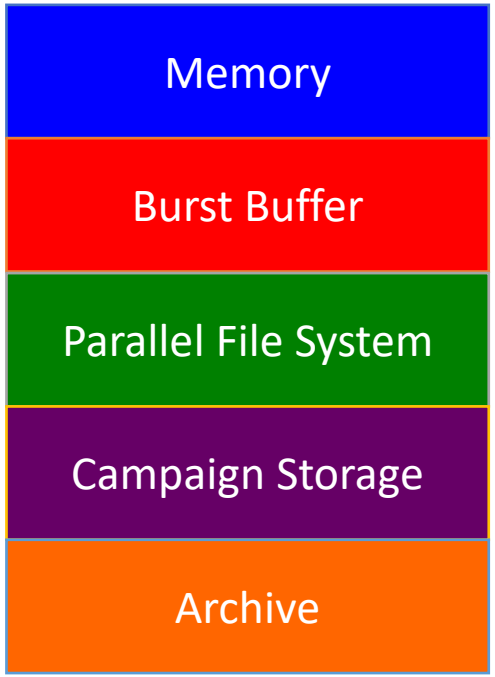
The now infamous fade out slide circa SC14

HPC Before 2016



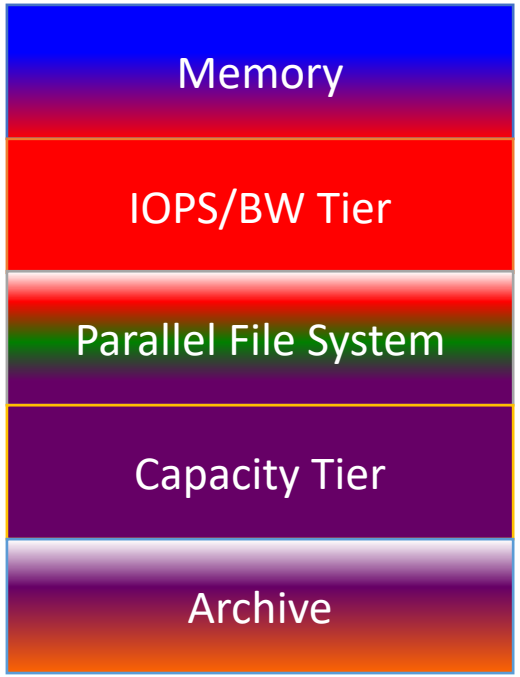
Sites that run PB working set problems for months to years needed this, others just added disk to PFS or Archive

HPC After 2016



- Economics
- Add BB (Flash)
- Add Campaign (low cost disk, slightly lower function than PFS)

HPC Post 2020/21



- Burst Buffer Software matures
- Campaign Storage leverages cheap dense, hard to write disks, extreme protection

Dropping ratio of Flash/Disk Cost for Capacity driven by cloud scale

- Allows larger in system storage gives rise to all flash file system startups (2+) and additions to existing pfs technology to enable all flash in system file system (erasure/etc.)
- Agile disks-the past, all drives hard to write and failure prone need far more erasure/attention to write (immutable/etc.)

Tape Archive for low BW, disconnected from power

IOPS/BW Tier software must mature due to size/durability increase (erasure)
Capacity tier disk dominated, tape only for lower BW/disconnected from power applications

Interesting observation: Just as we did early in the life of checkpoint-restart, we bought capacity and got BW for free and now we have to be keenly aware of the BW and Capacity concurrently,
there seems to be a similar observation in EOD/EOS but maybe its worse. You always had to worry about both Capacity and BW in EOD/EOS, but you hear interesting information like: It would be best if we could reprocess all the data from the beginning of the instrument/experiment(s) life every month - Oh My .
*** Just like in checkpoint – we need to follow network/storage economies to afford a solution!

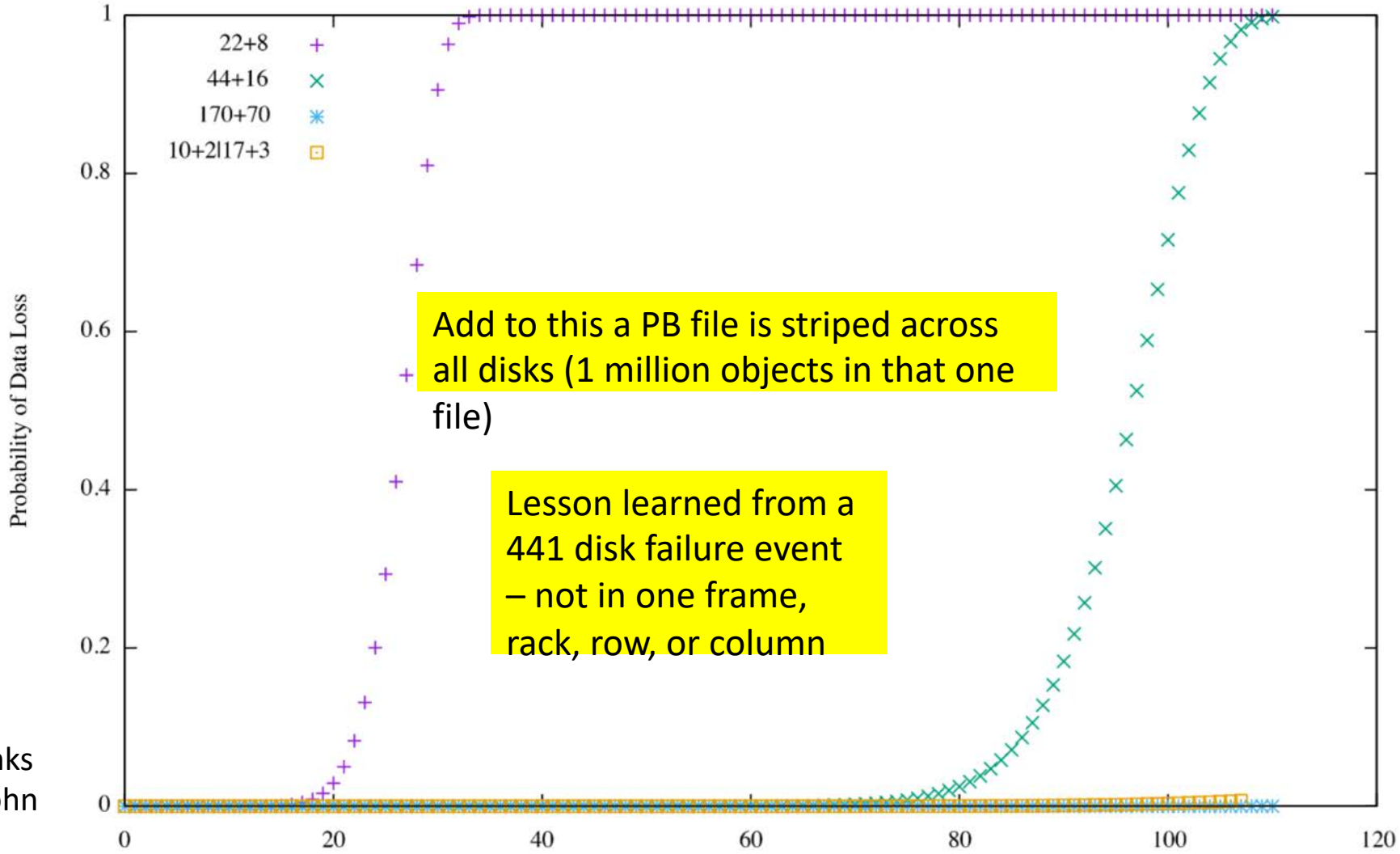
Changes in the number if tiers will occur due to economics/technology from time to time.

For each tier you need the appropriate speed, size, and protection

Add to that the very large dataset problem (lose one byte and you lose a petabyte) how much is enough erasure
(not just time to first byte lost but also rate of loss)

Add to that tape tech will be cheap and less reliable (due to cloud drivers), will need a much more elegant erasure for that too

You heard about data explosion: If we use disk for most of our capacity, how much protection is enough?



Add to this a PB file is striped across all disks (1 million objects in that one file)

Lesson learned from a 441 disk failure event – not in one frame, rack, row, or column

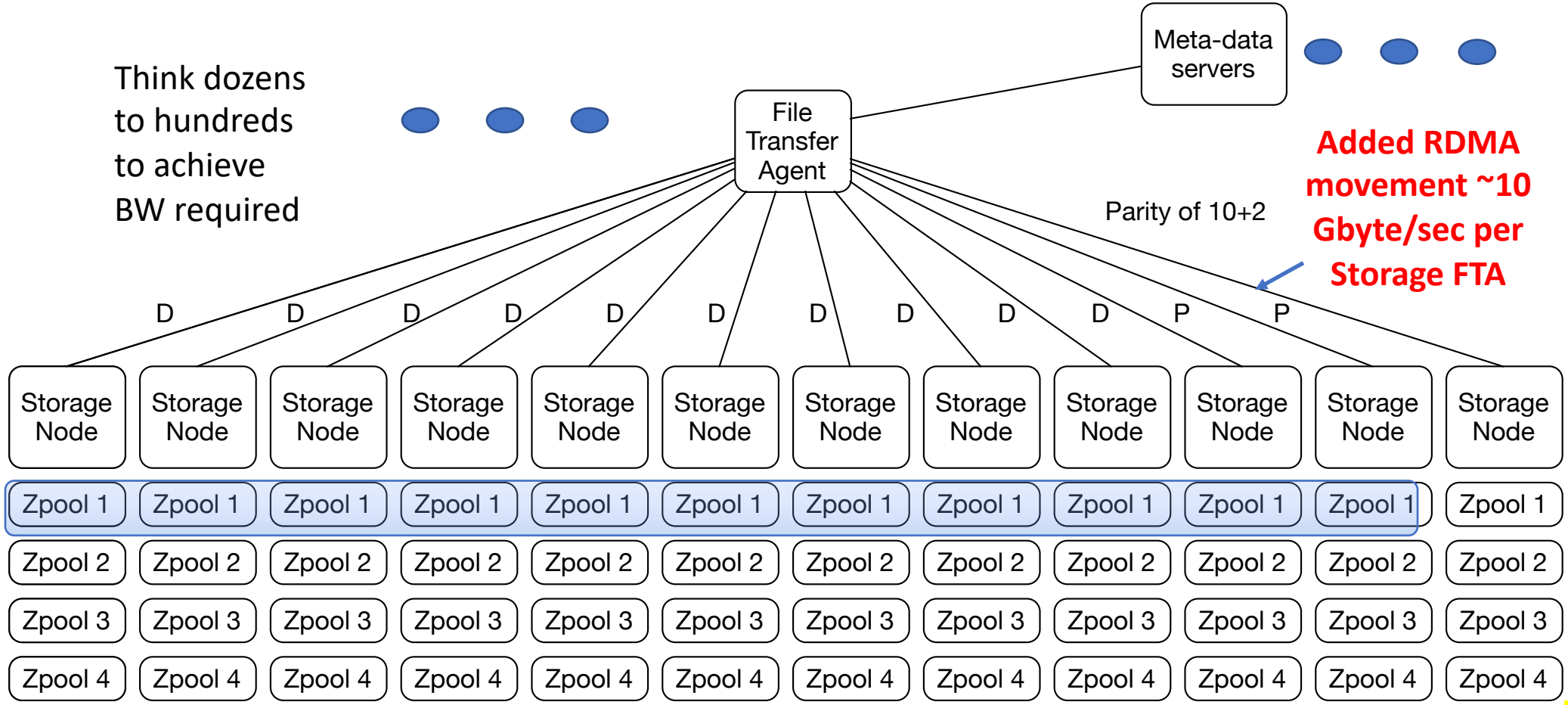
Thanks to John Bent DDN

Data Loss Probabilities with One Trillion Objects

Infant loss, random loss, spatial correlated loss, aspatial correlated loss

Its not just time to first byte lost, its also if you lose a byte how much did it take with it

Overcoming Correlated Failures
Multi-tier erasure is working for now

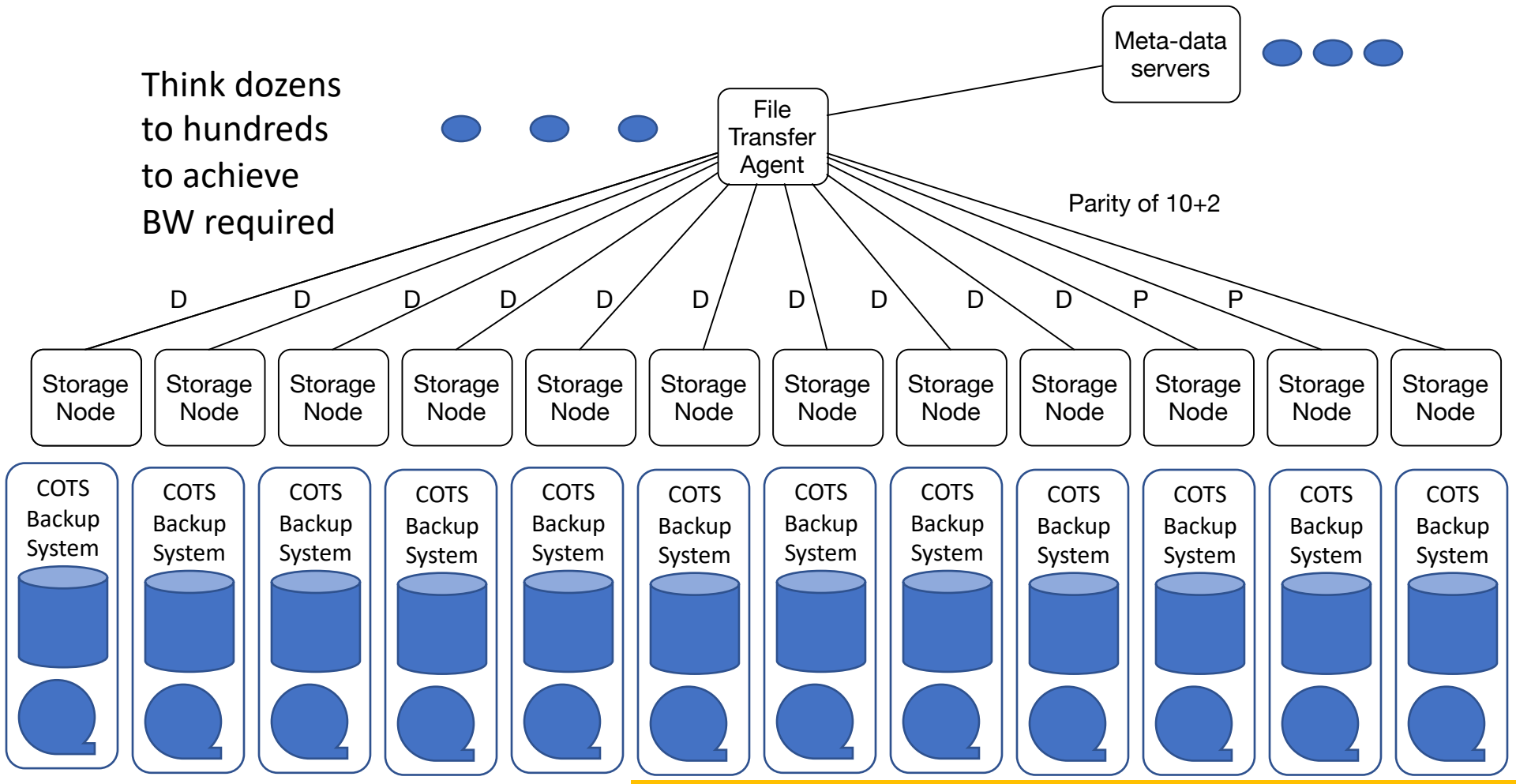


For random failure you want fast rebuild which needs more failure domains, but for correlated failures you want to limit failure domains

Are there better solutions?

MarFS Campiagn Store in Production 2 yrs

Erasure on power managed disk or tape for extremely resilient archives, being prototyped



The largest growing consumer of tape is cloud, they will insist on cheap and just protection. Expensive tape technology is likely dead.

We will very likely have to follow the cloud folks and use erasure, but a much more flexible and scalable solution than any of the current or past RAIT solutions.

A pretty easy extension to MarFS

Is there a better way?

Peter Braam says MarFS is the best HSM backend and we don't even realize it
2 tier erasure, parallel async hole poking, doesn't require a separate db/namespace to reconcile, etc., interesting complement to someone that is not fond of HSM's

Wont that new storage class memory save us?

- Lots of headroom in flash tech and not much in disk/tape tech, so lots of bytes of storage purchasing power at play for tech capable of volume.
- Volume production is extremely expensive and will be leveraged heavily once invested in making adoption of newer tech slower than it might be
- New tech going after low end memory market where margin can be high
- Some are calling nvm part of their memory size and while that might be true for some simpler use cases, for complex simulation where I come from that's a pretty laughable concept. DDR is greater than an order of magnitude too slow so nvm used as active working set is wishful. Your mileage WILL vary.
- New tech will have some uses but by 2025 but as a replacement for working set memory or higher capacity flash is not likely at least for all uses.
- We do need access methods (byte addressable) that allow specification of non volatile and performance expectations, but frankly just a nice non block (variable length interface (object get/put stream or KVS) to lots of flash would enable lots of innovation.

Namespace(s) One is appealing but maybe optimistic

- Every time tiering of storage comes up there is a clamor for a single name space because it looks elegant and simple but:
  - Over time, namespaces have had performance/functionality expectations attached to them
  - The classical grep from hell problem on hsm's still exists, users think because it's a mounted file system it should act like their conception of a mounted file system, if it doesn't act that way they hit control-C and run it again and again.
  - Single name spaces look like they work well for instrument/machine driven workflow but maybe not so well for human driven exploration.
  - The current fascination with workflow automation and machine learning pipelines are driving this discussion far to one edge, remember balance Daniel-san!
- Maybe instead of lunging at one name space notions we should invent and push a new concept in how namespaces are conceived?
  - Example: current mount options are RW and RO but that doesn't really serve us well

  Should there be md R, md W, data R, data W, data append only, data version write only

  These concepts are far more useful for leveraging namespaces be they flat or graph

  Maybe IOPS Tier mdRW,dataRW, campaign mdRW,dataR, archive mdRW,datanorw
- And more metadata around names that is searchable, kind of like the Grand Unified File Index (LANL's version of this)

Access methods – wins seem likely and timely

- POSIX access method has served us at least in some ways well and people are becoming use to loosening POSIX semantics, especially related to particular namespaces (see namespace slide)
- We have Checkpoint/Restart on the run, adding value to the data for better data mgmt./use makes sense. EOS/EOD is similar and growing. Adding value has been a 2nd class citizen in storage systems (relegated to living inside a byte stream/file), but long ago different file types were the norm (KSDS (KVS), ESDS (Log), etc.). Do file systems need to treat added value as first class citizens?
- User space, right sized, composable, discoverable, data mgmt. services all seem like they have been prototyped but not in service yet. There is work to move this from research to production for sure.
- Leveraging lower latency/high bw storage and byte addressable may/will require new access methods
- New access methods may loosen the bridle on the storage vendors thinking they have to do blocks
- New access methods may enable compute in network/storage beyond todays simple examples (compress/erasure/etc.) instead things like (multi-dimensional, etc.)
- Problems:
  - too many access methods will confuse and disincentivize enabling innovation, need to find a few common powerful ones to push
  - Too high of a level and it wont help much
  - Too low of a level and it will be a small niche and not generally helpful to the overall community
  - Too complex and it wont be adopted widely
  - Too simple and it has little value

Freedom from Tyrannies

- Continue to loosen up POSIX (it just takes time and effort)
- New Access Methods to incentivize innovation enabling activities
  - Instead of going after POSIX as the only chain, go after other chains like Block. Blocks make the IO stacks thick as to file systems. Think about if disks and flash and OS's could drop block support. Innovation near the hardware is shackled to blocks. That doesn't mean byte addressable everything, just some variable length methods (KVS, Object get/put streams, etc.) would be useful.
- New devices (I know its too expensive, but is it?) (recall you need more BW per byte than ever given interesting EOD/EOS mentioned before)
  - Why has tape served the world for so long? Well its extremely cheap, but partly because its multi-dimensional density but partially because its more linear than square so density improvements and bandwidth improvements scale reasonably together.
  - Disks do not have this going for them, but you can parallelize them cheaply.
  - Are their other technologies that give you what tape had and are inexpensively parallelizable??
  - New applications coming on line have scary bandwidth characteristic (need to re-process from the beginning each year).

Trains leaving the station that might help

- Byte addressable NV Storage – just because economically it doesn't look great right now, stay on top of it as economics change
- NVME/NVMEOF – fastest growing storage sector I have ever seen
  - Extremely well supported RDMA local and remote access to block devices
  - Ubiquitous global accessibility
  - Leverages interconnects
  - User Space Drivers exist and are well supported
  - Samsung pushing KVS onNVMEOF – implies variable length capability, may be our opening to push variable length capability to free disk/nv storage vendors from block tyranny
  - Growing faction of vendors wanting to put compute elements as addressable elements in the NVME/NVMEOF environment, compressor, erasure, encode, etc. May be our opening to get our much and long wanted intelligence in network/storage including indexing
  - Ubiquity may enable right sizing/composability/buzzword bingo stuff
  - MASSIVE TRAIN that gets bigger and more powerful by the day