

SDC 18

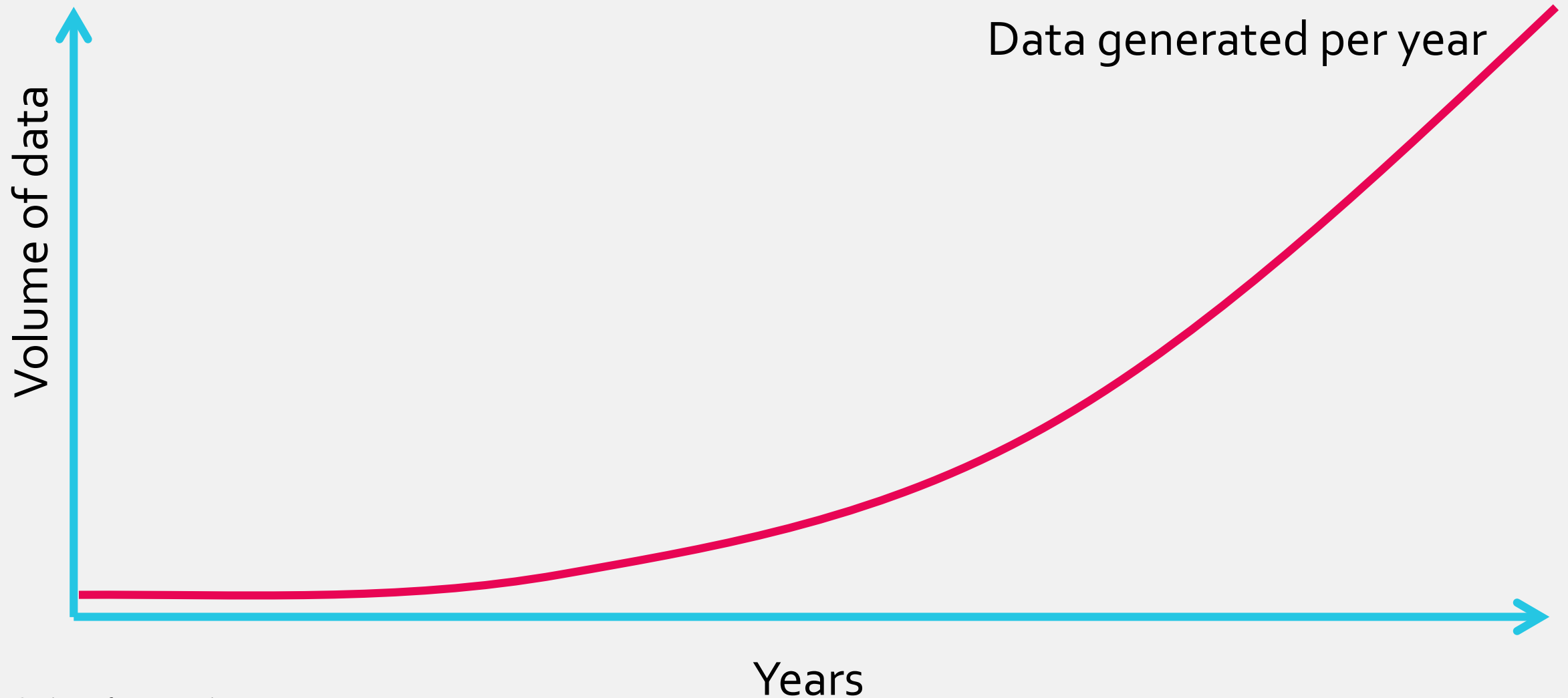
September 24-27, 2018
Santa Clara, CA



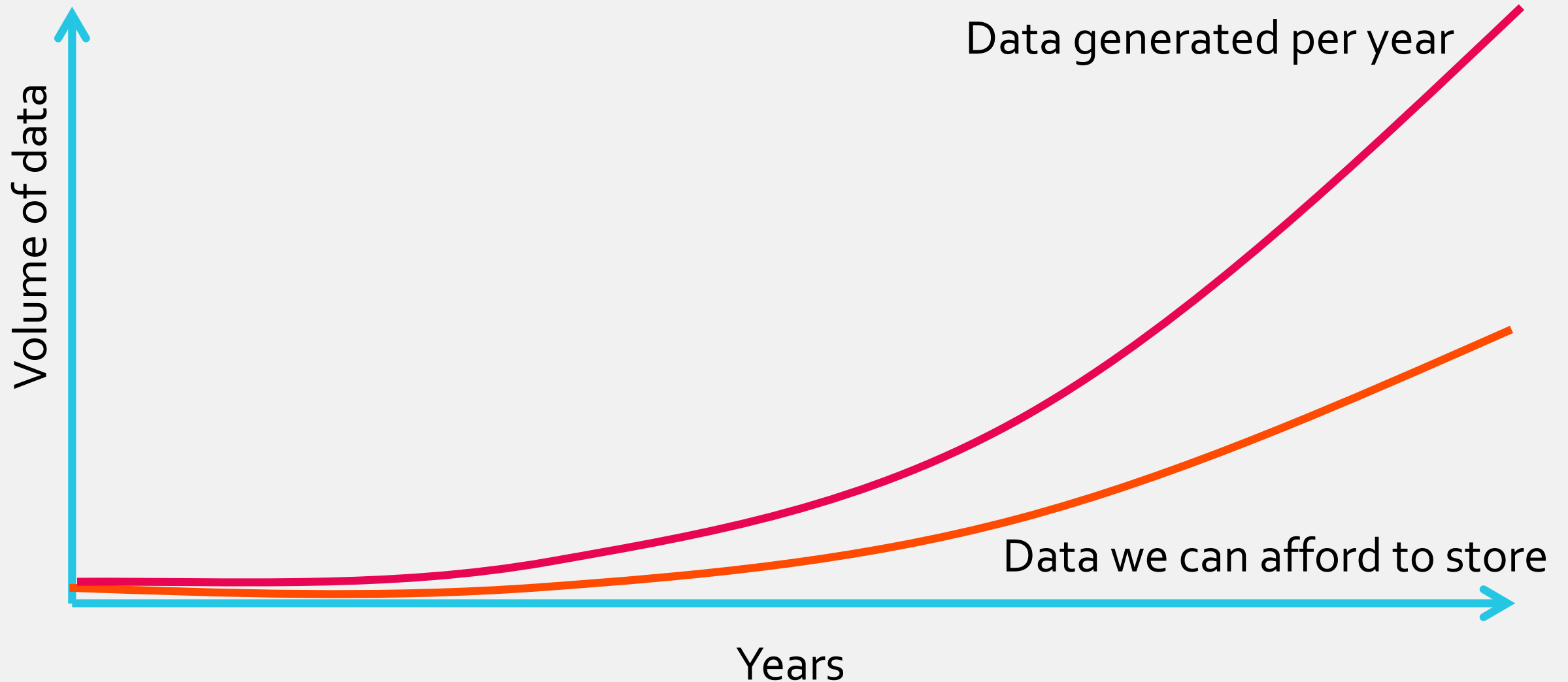
Glass: A New Media for a New Era

Austin Donnelly
Microsoft

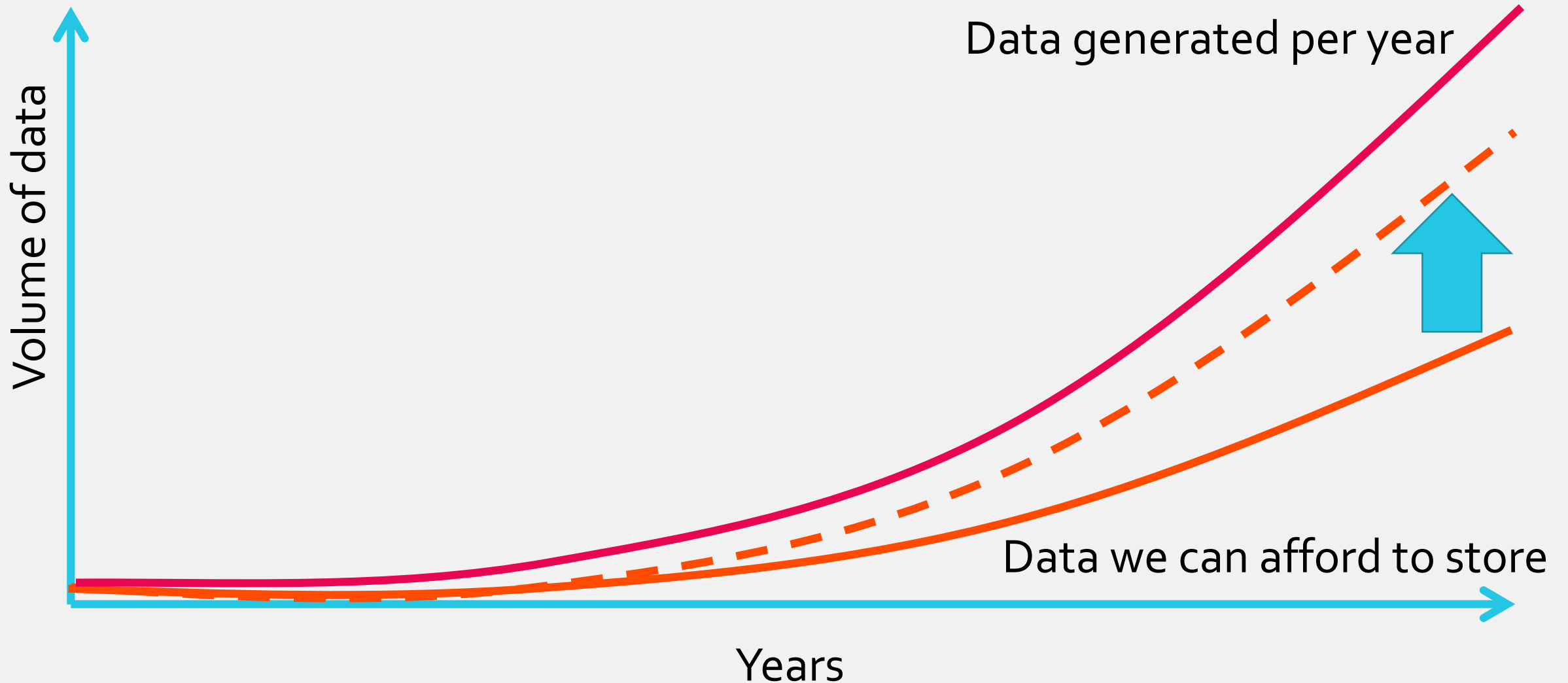
Data volume is growing.....



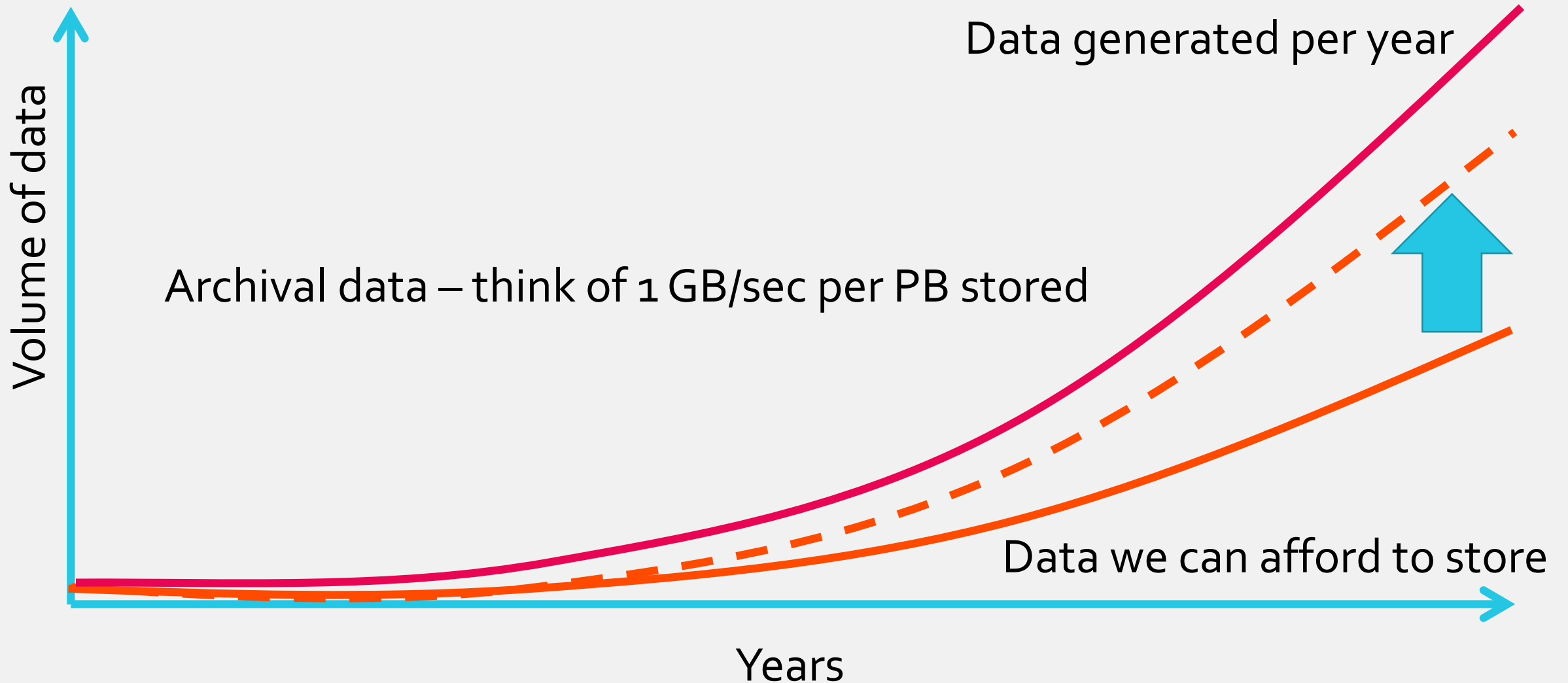
Data volume is growing.....but storage is \$\$\$



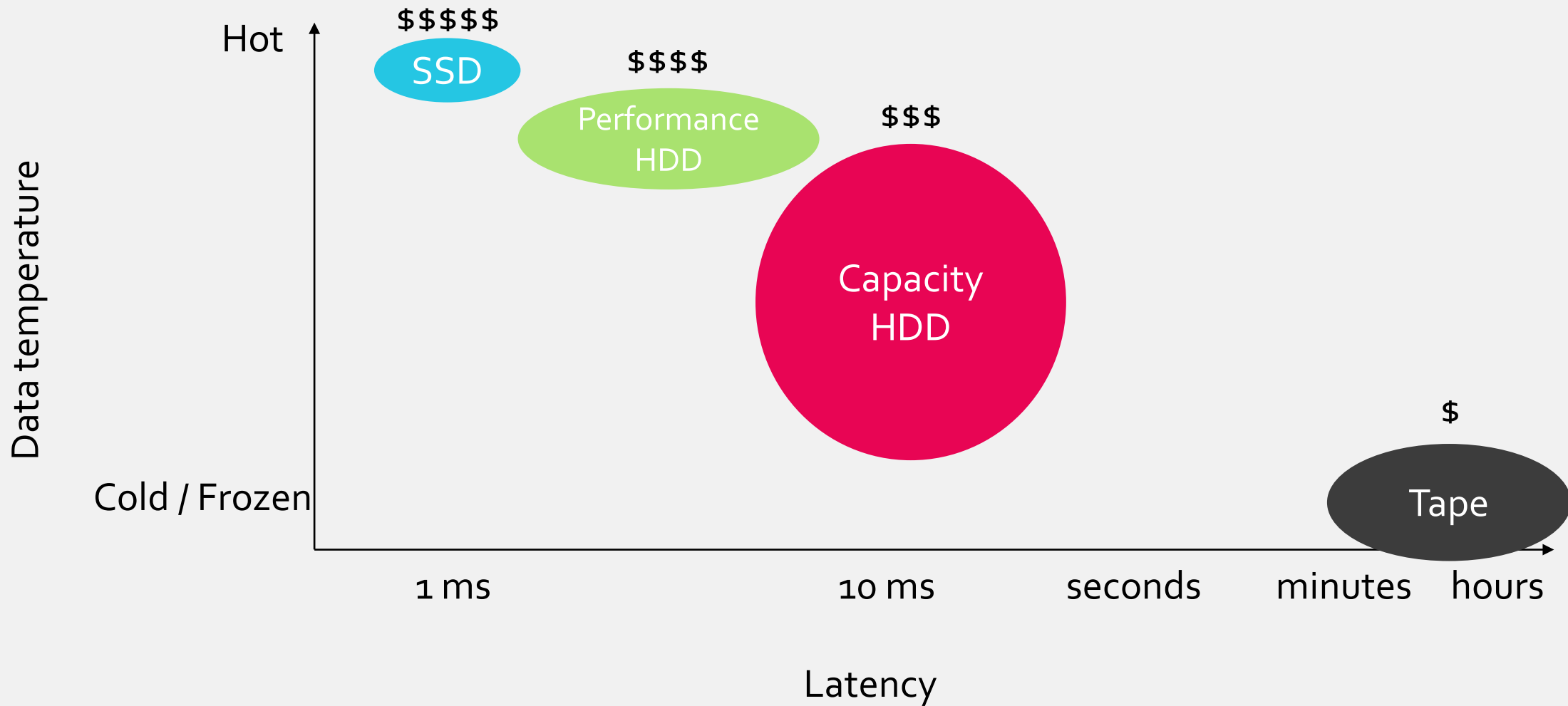
What we are trying to do....lower \$/GB



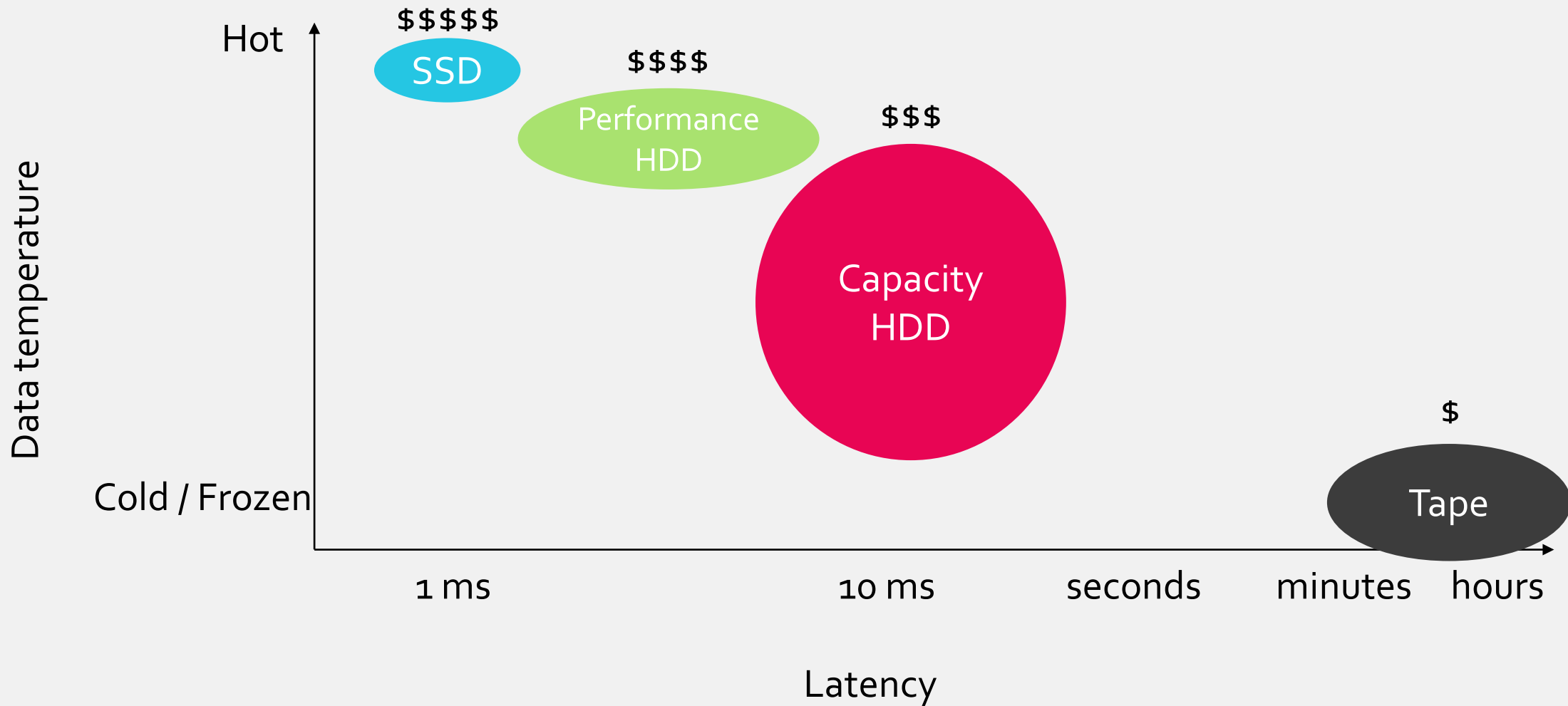
What we are trying to do....lower \$/GB



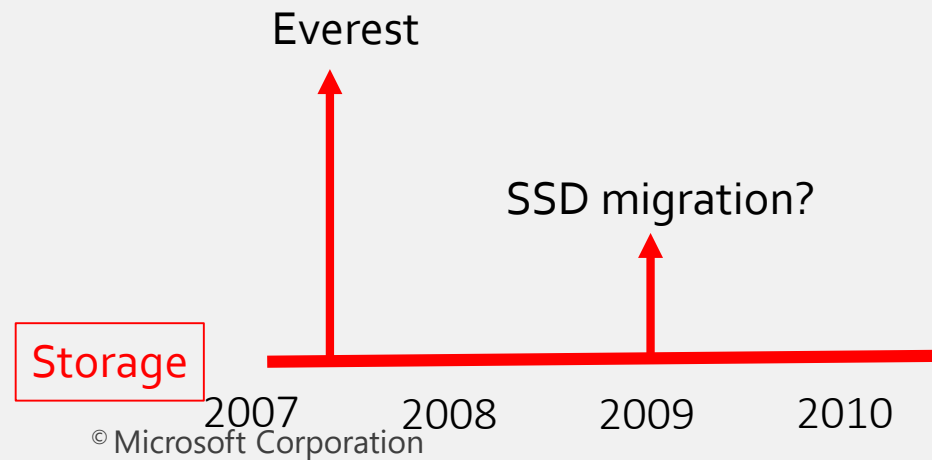
Existing technologies...



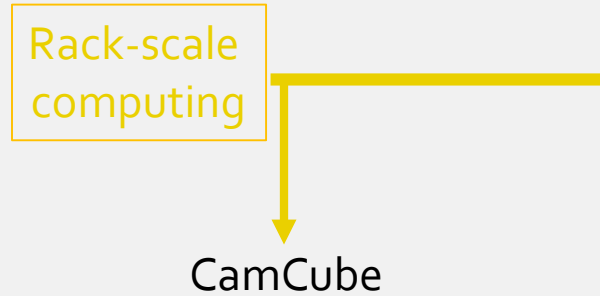
Cool storage “latency gap”



The journey from 2007 to a new media.....



The journey from 2007 to a new media.....



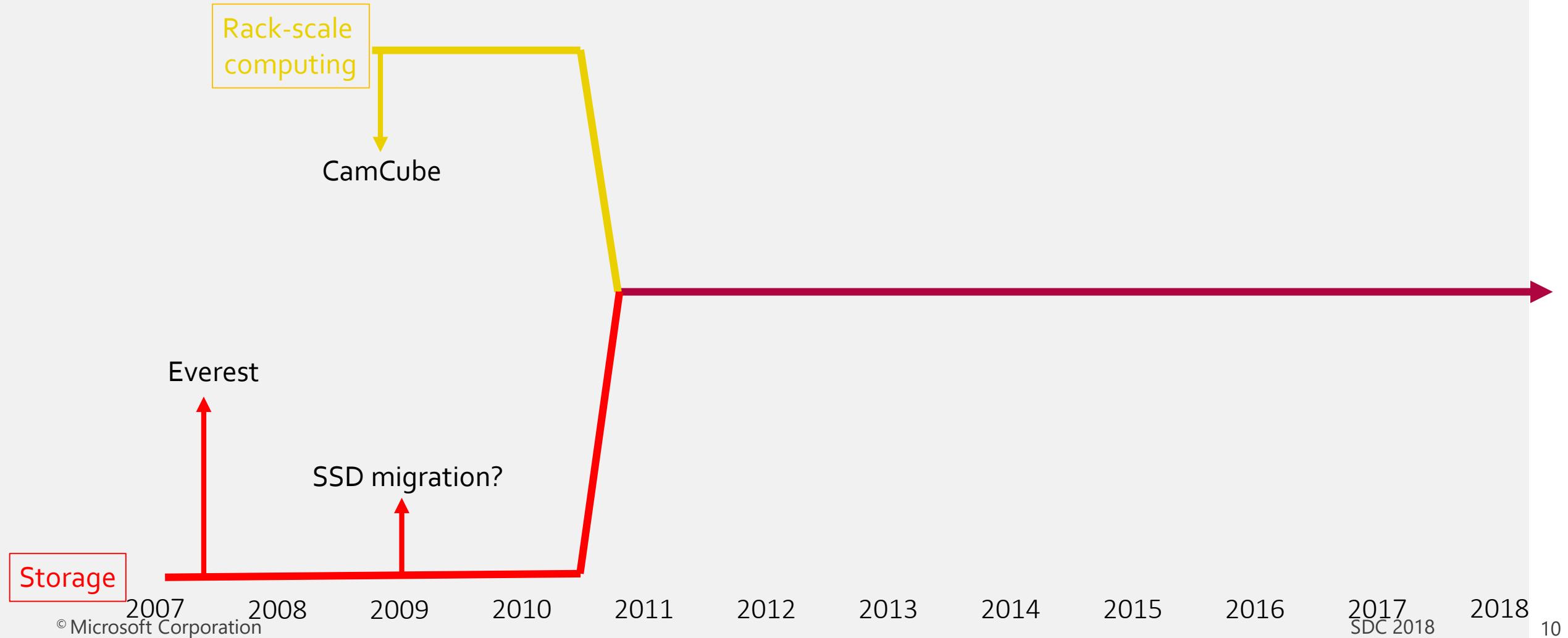
Everest

SSD migration?

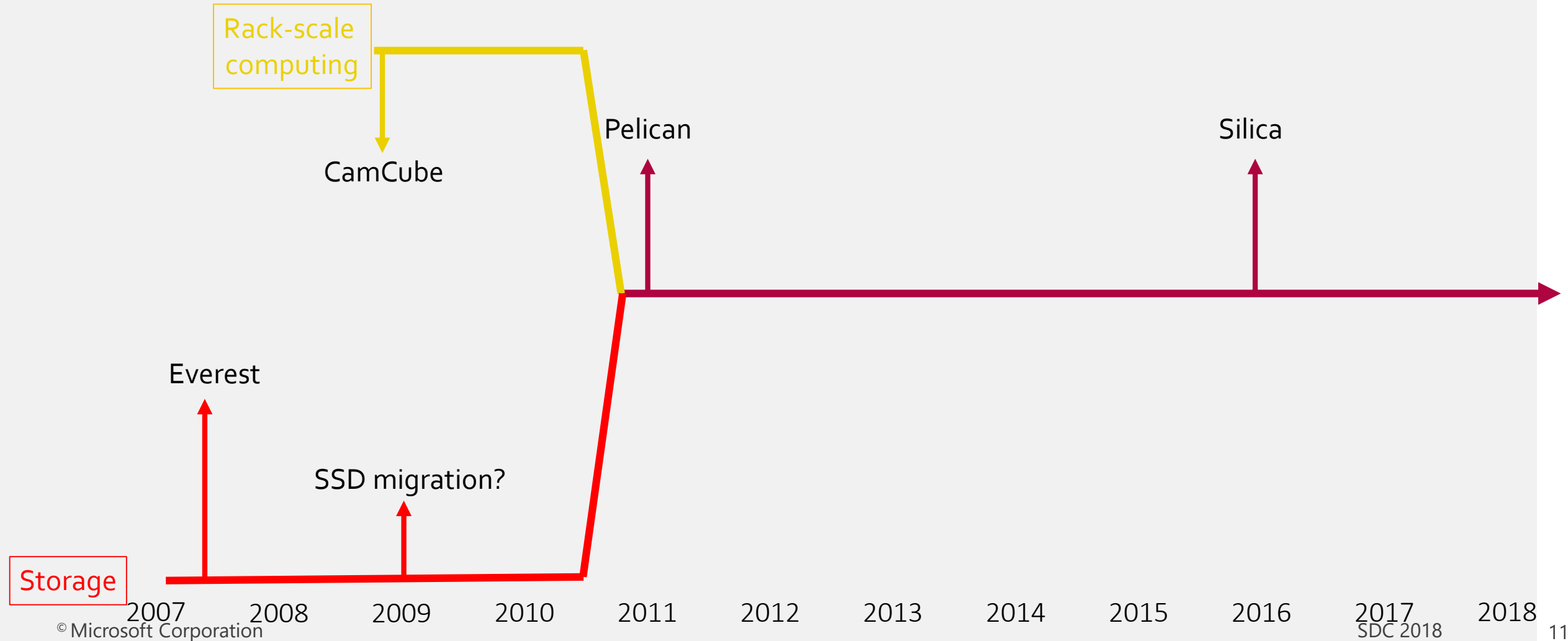
Storage

2007 2008 2009 2010
© Microsoft Corporation

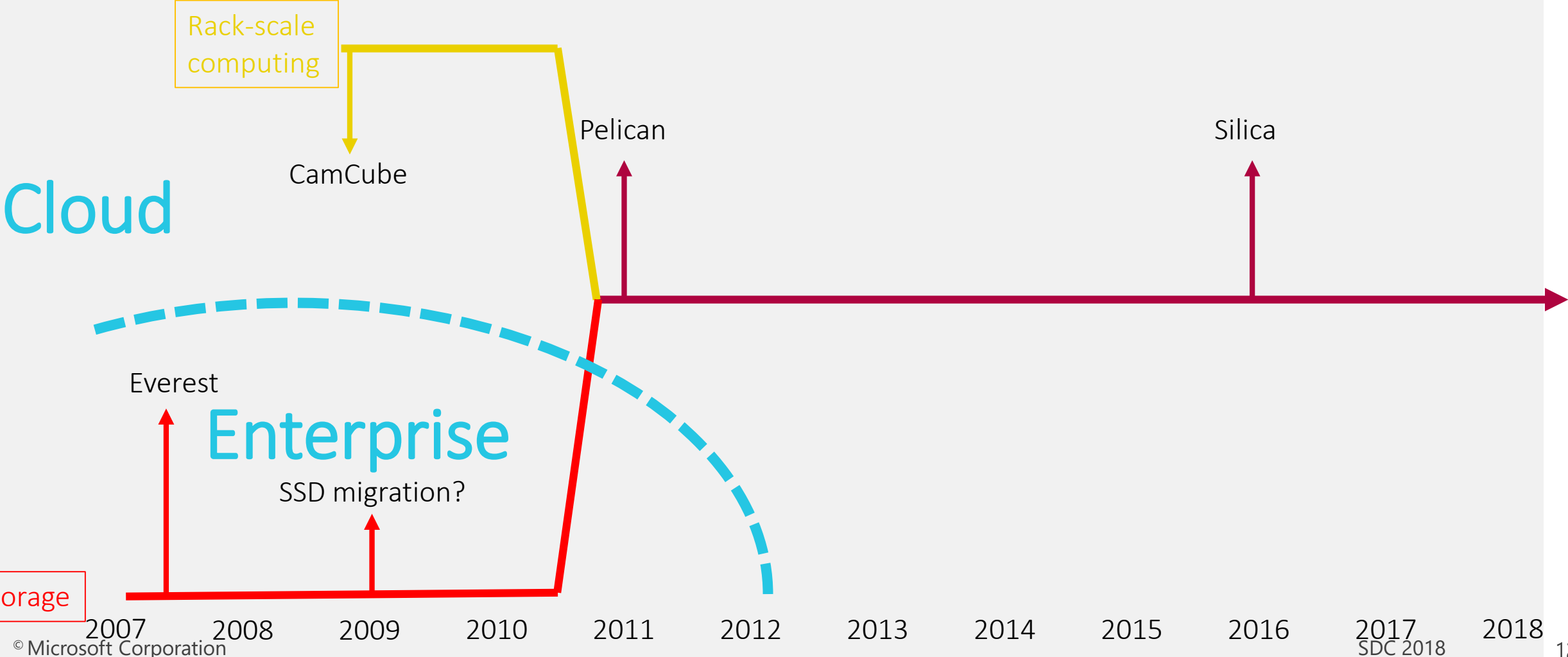
The journey from 2007 to a new media.....



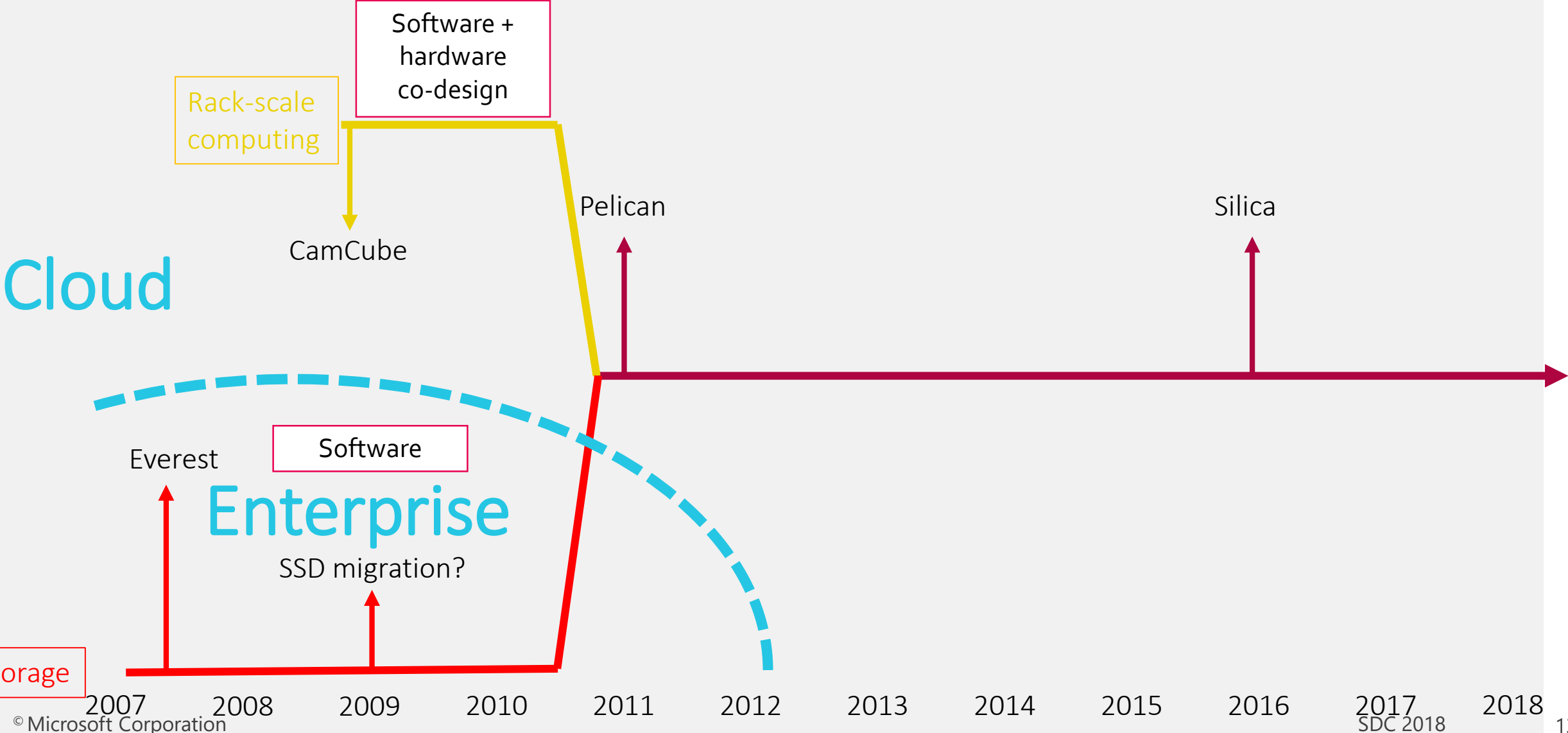
The journey from 2007 to a new media.....



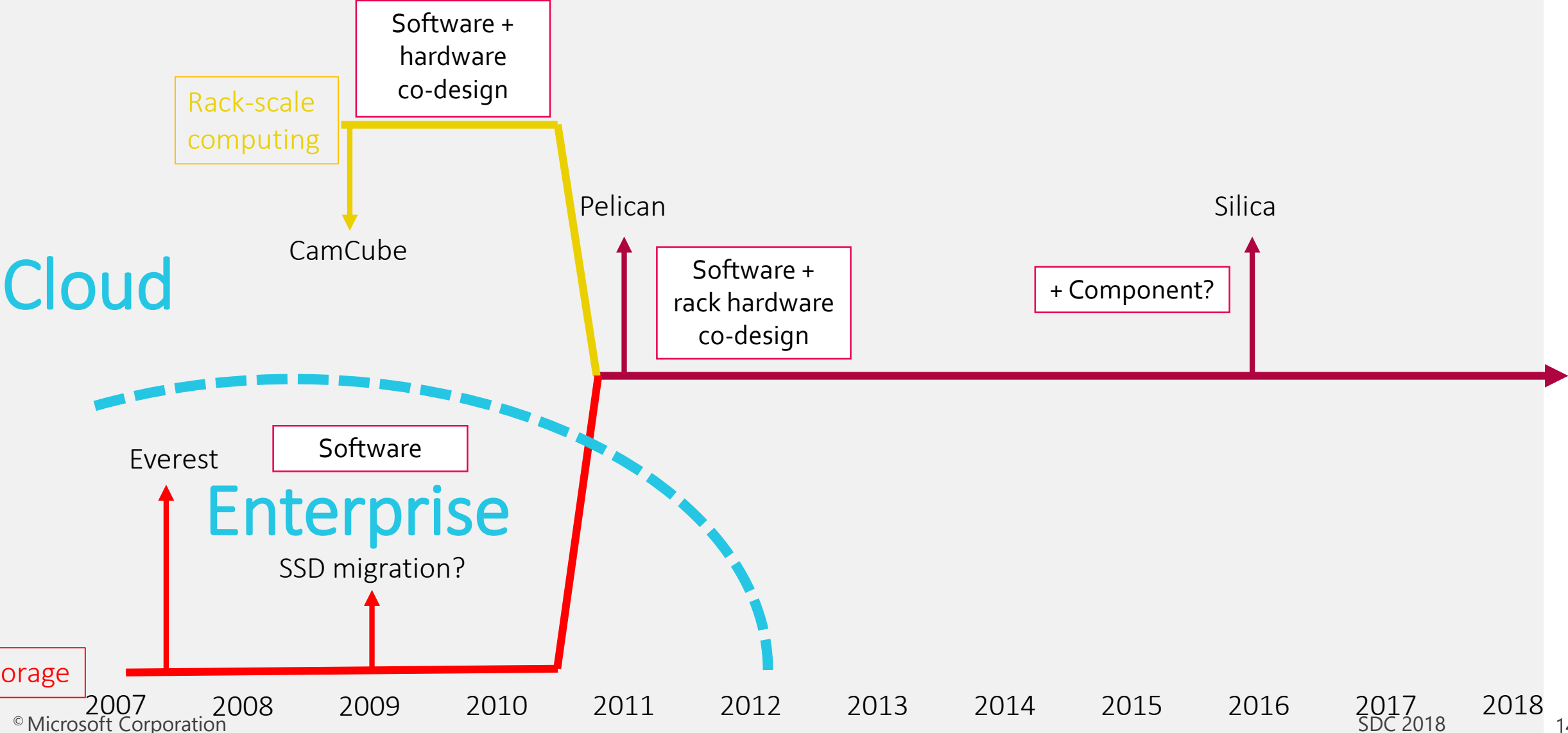
The changing thinking.....



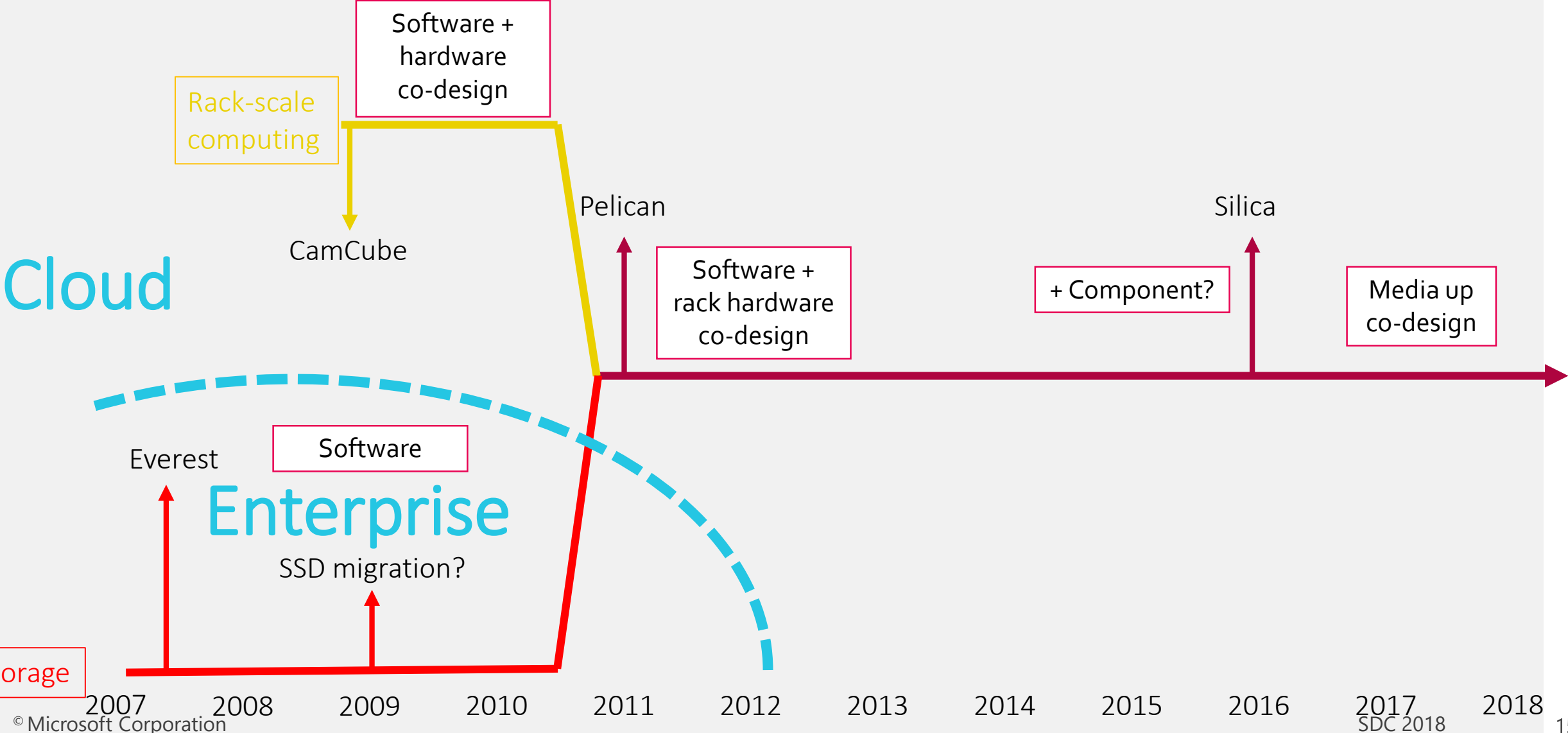
The changing thinking.....



The changing thinking.....

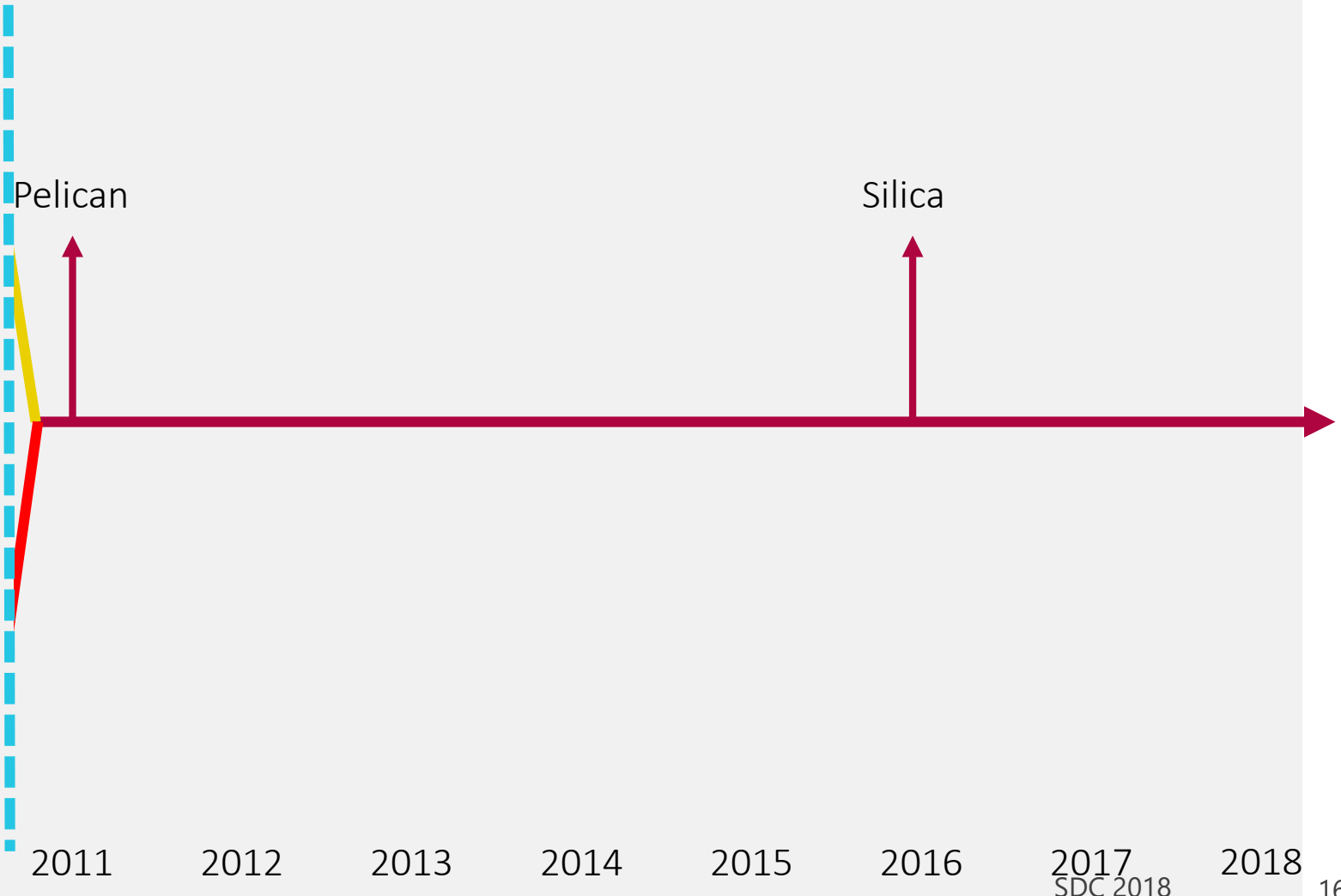


The changing thinking.....



Today's talk

Glass: A New Media for a New Era

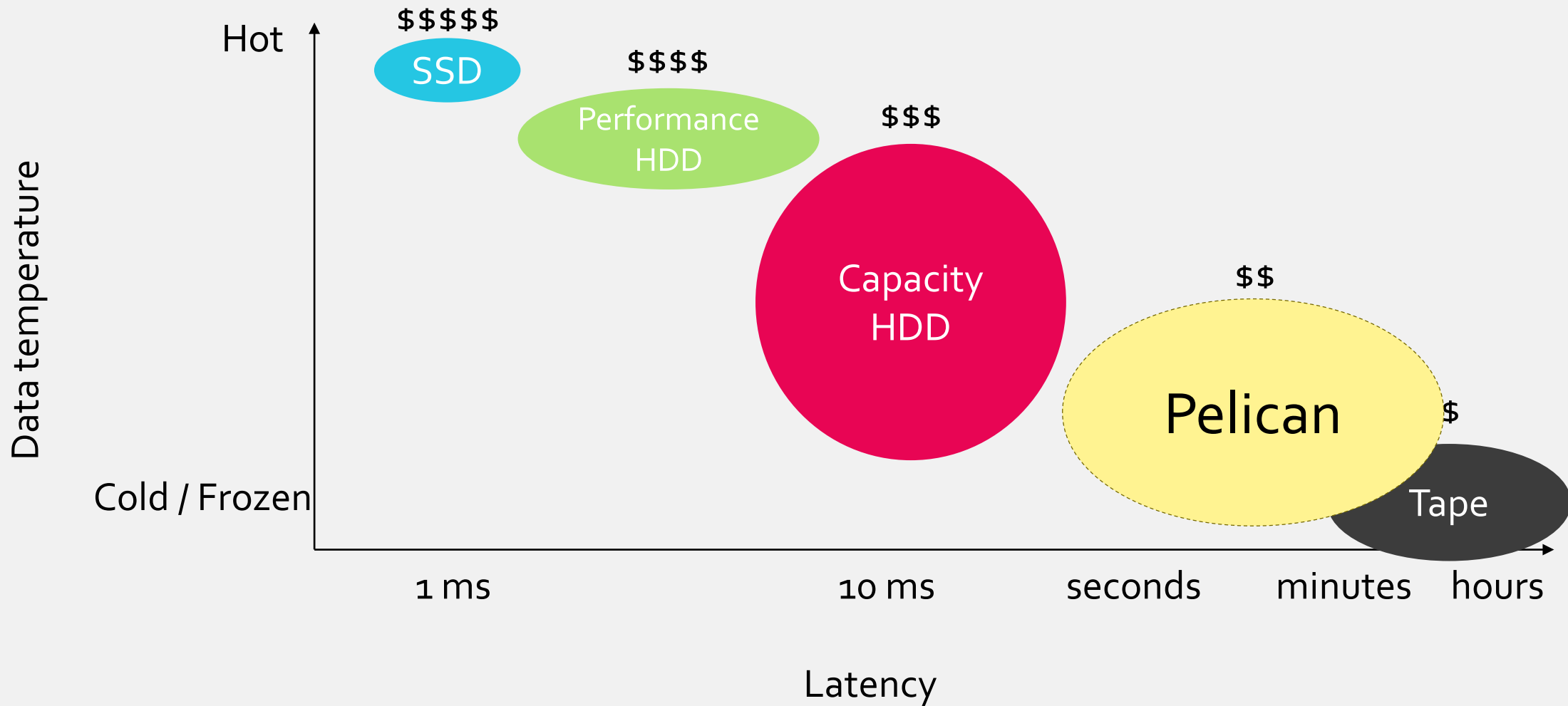




Pelican

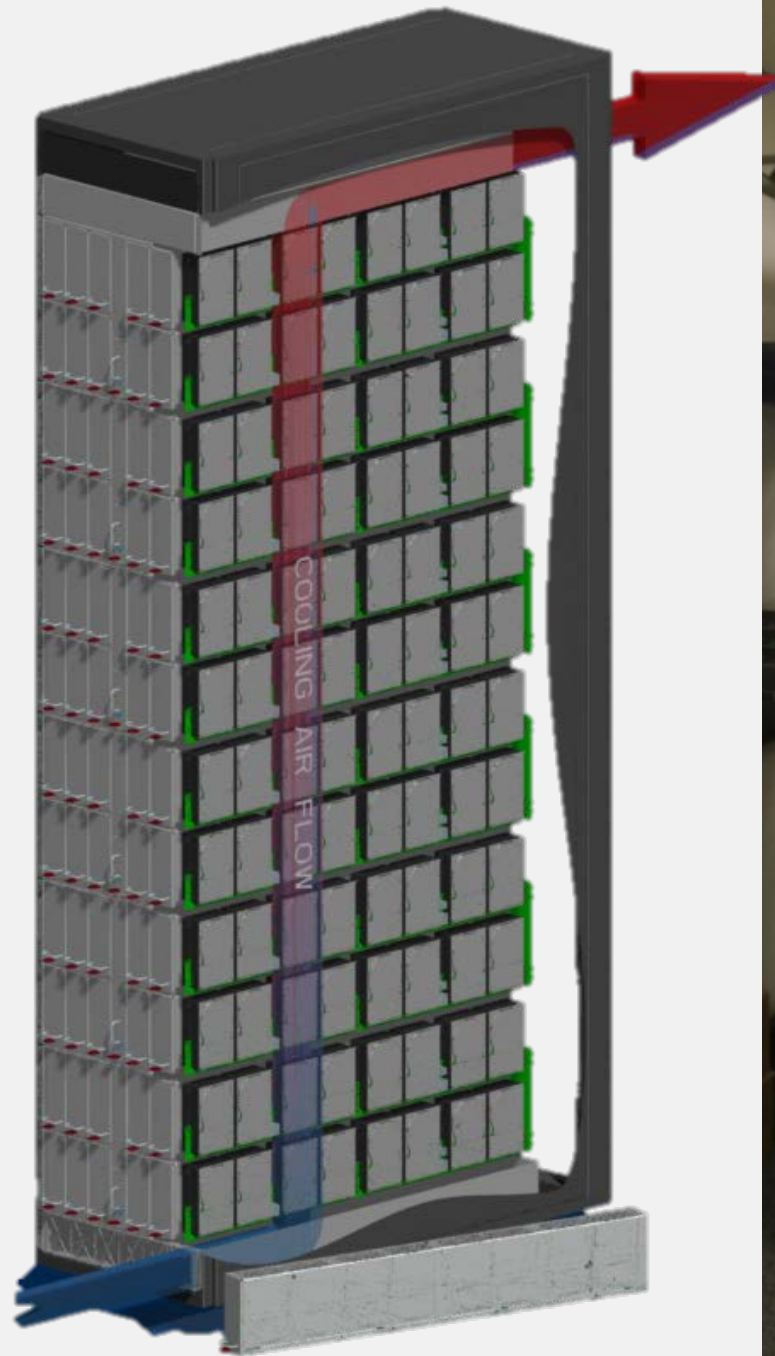
Low-cost HDD-based active archive

Cool storage “latency gap”



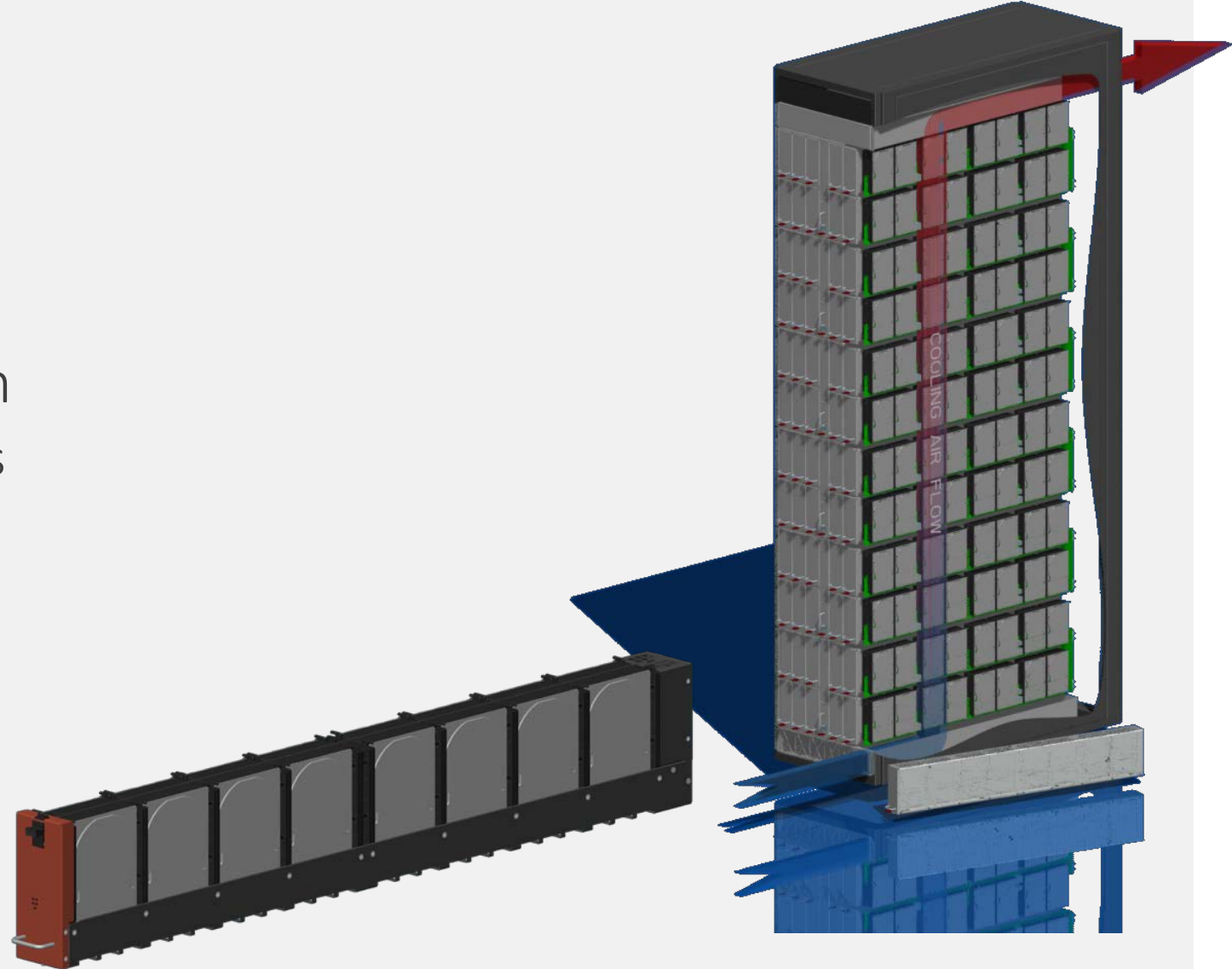
Pelican

- 2 Servers, PCIe rack wide
- 1152 SATA 3.5" HDDs
- Vertical Cooling
- Resource constraints
 - 1 disk/column
 - 2 disks/tray
- Manage in software



Pelican: Interesting design choices

- HDDs spun down
 - How do the drives cope?
 - How does this impact latency?
- Disaggregated rack-scale design
 - Disks can migrate between servers



Archive drives

- New class of HDDs
- Optimised for minimum \$/GB
- Targeting cold workloads:

“The WD Ae hard drive is best suited for cold storage, backup and data archiving where data is stored on disk but rarely if almost never read again”

-- WD6001F4PZ1 datasheet

- Workload is quantified as TB/year
- Lifetime affected by:
 - POH
 - TB transferred
 - Spindown cycles

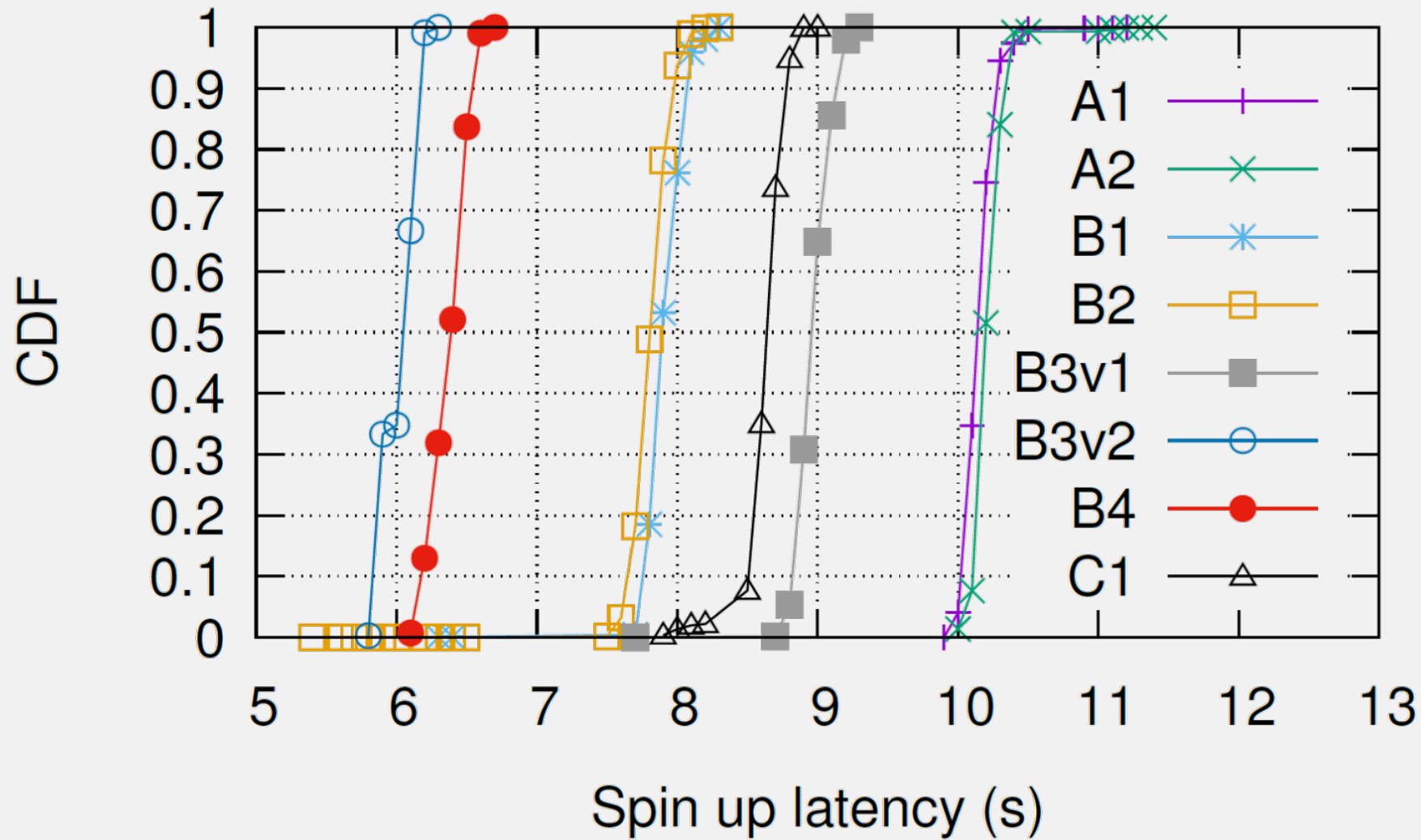


Drive line-up

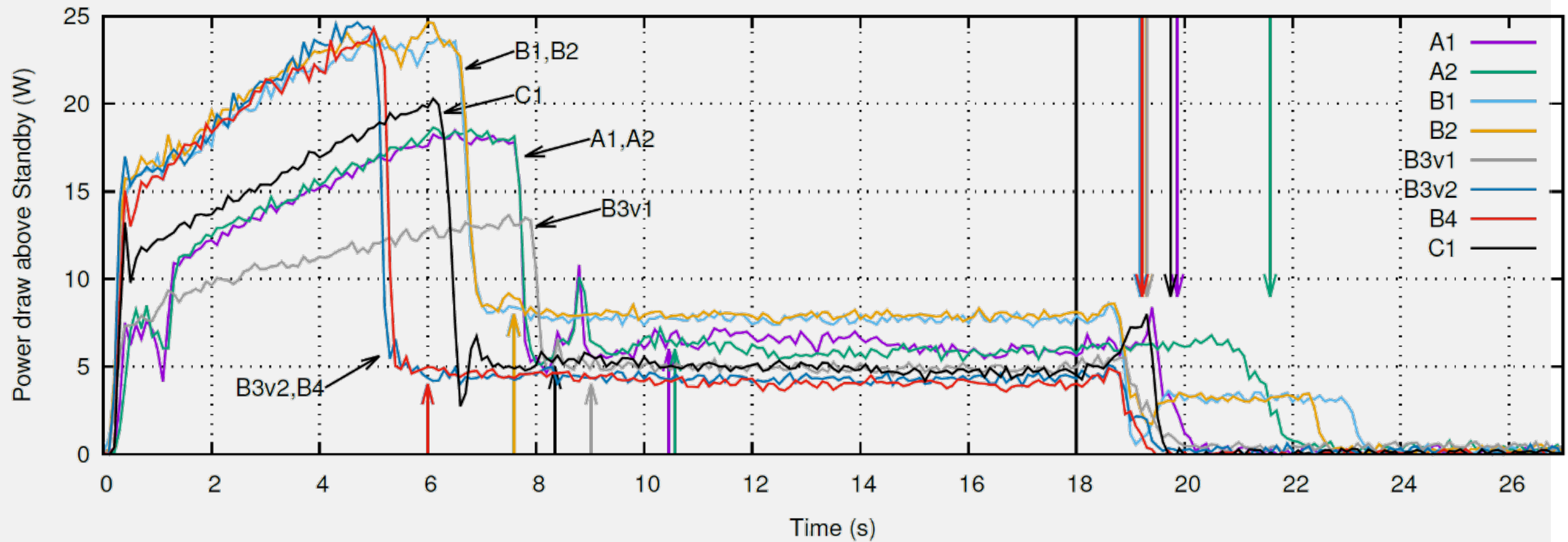
Name	Technology	Spin up (s)	Capacity (TB)
A1	Auto SMR	10.1	8.0
A2	HA SMR	10.2	8.0
B1	PMR	7.9	4.6
B2	PMR	7.8	4.5
B3v1	PMR	9	4.9
B3v2	PMR	6	4.9
B4	PMR	6.4	6.1
C1	Auto SMR (?)	8.6	8.0

SMR = Shingled Magnetic Recording PMR = Perpendicular Magnetic Recording

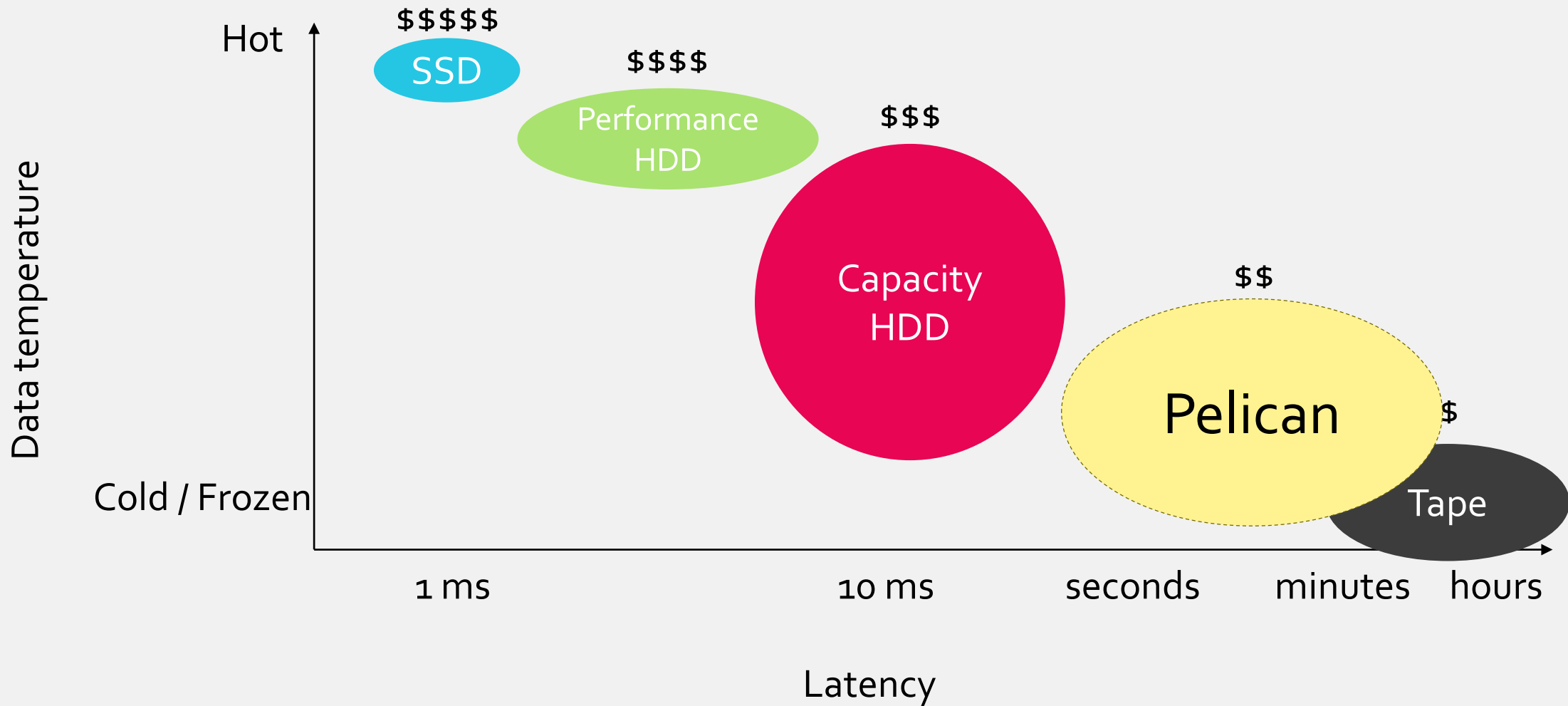
Spinup latency



Power draw

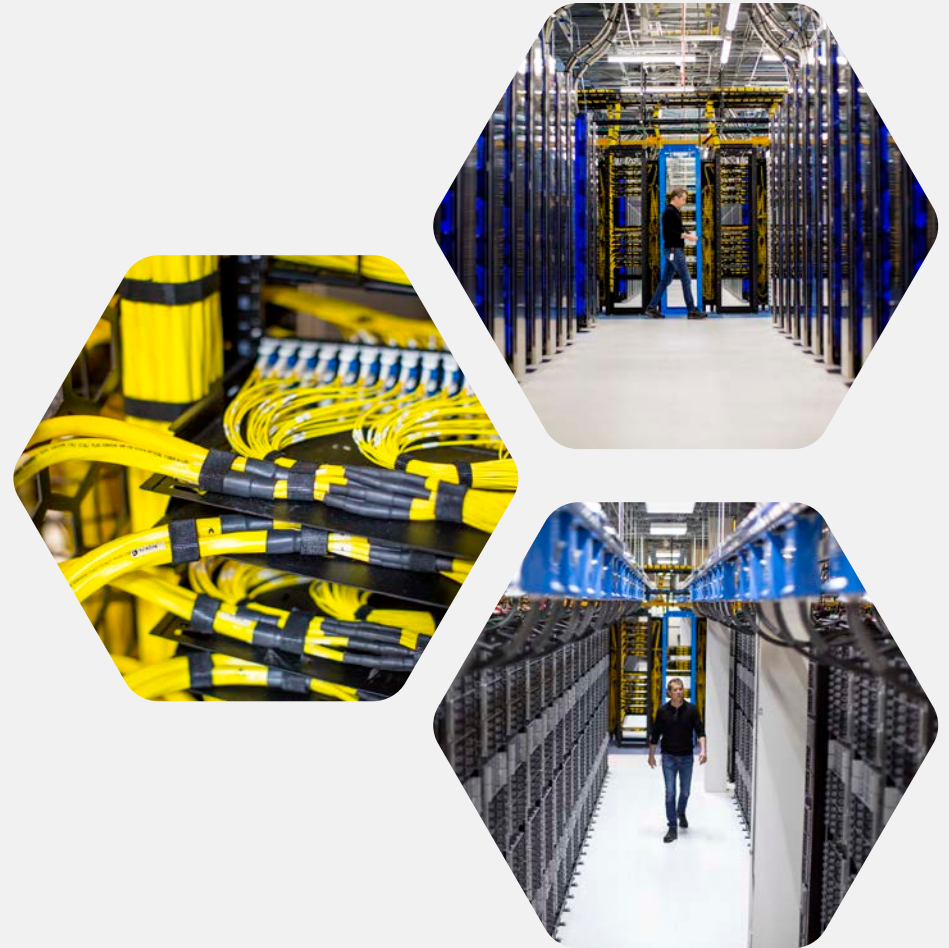


Cool storage “latency gap”

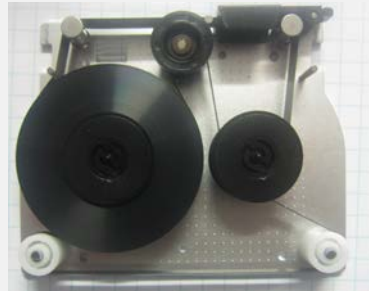
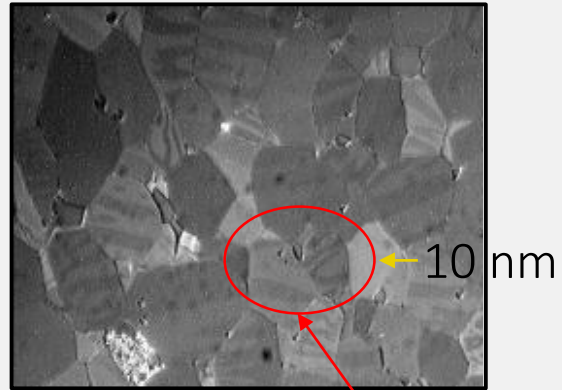


Pelican summary

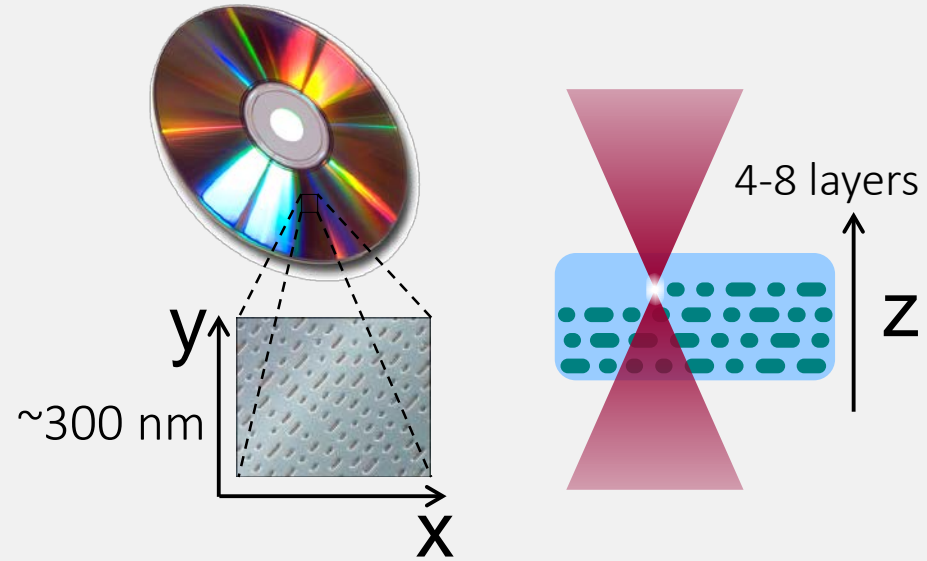
- Archive drives are effective, provided workload is managed
- Spindowns do **not** seem to affect drive reliability
- And they are probably the cheapest \$/GB HDD-based storage...
- We've tuned everything to minimize \$/GB
 - We've built highly efficient HDD storage for cold data.
 - We've worked with the HDD vendors to drive costs down.
- Let's go back to basics...



Long-term storage media



Magnetic storage



Optical Storage

Today's storage technologies

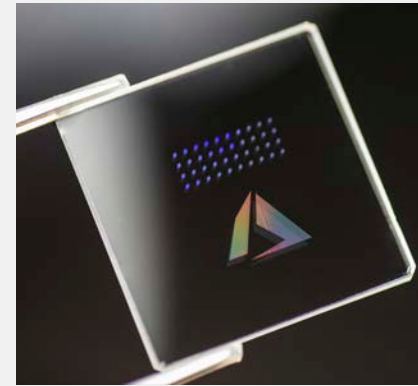
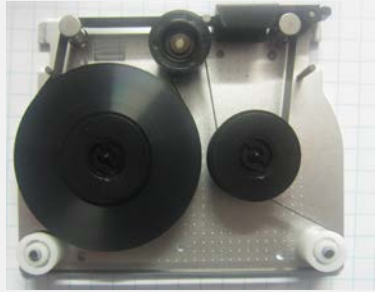
No storage technology deployed today in the cloud has been designed from the media up to support just the cloud

- HDD
- Tape
- Optical Disks (Blu-ray)
- Flash

Are there better media?



New media for a new era...



Magnetic storage

Optical Storage



Silica

Long-term active archive

Glass (fused silica/quartz)

- Storing data changes the structure of the glass
 - Write Once Read Many (WORM) technology
 - Archival
 - Persistent (think millions of years)
 - EMF-proof
 - No bit rot / disc rot (no scrubbing)
 - Cheap media
 - Seekable
 - Leave data in-place



Cuneiform tablet recording the allocation of beer, 3100-3000 BC.
© Trustees of the British Museum.

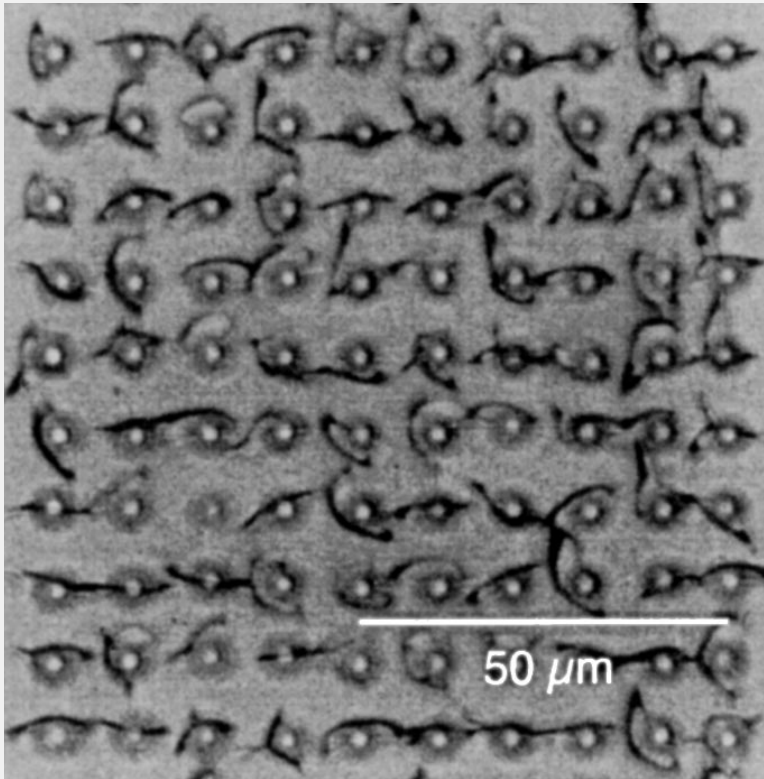
- Opportunity to really design from the ground up: think differently

Disaggregating write, media storage, read

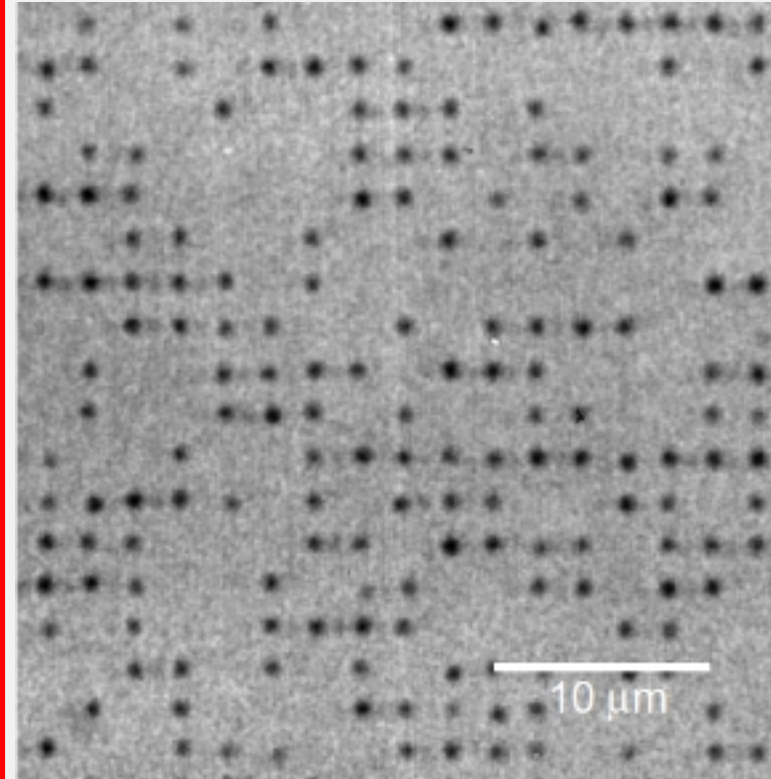


Optical storage using femtosecond lasers

Picosecond (10×10^{-12} s) laser
induces voids *with external stress*



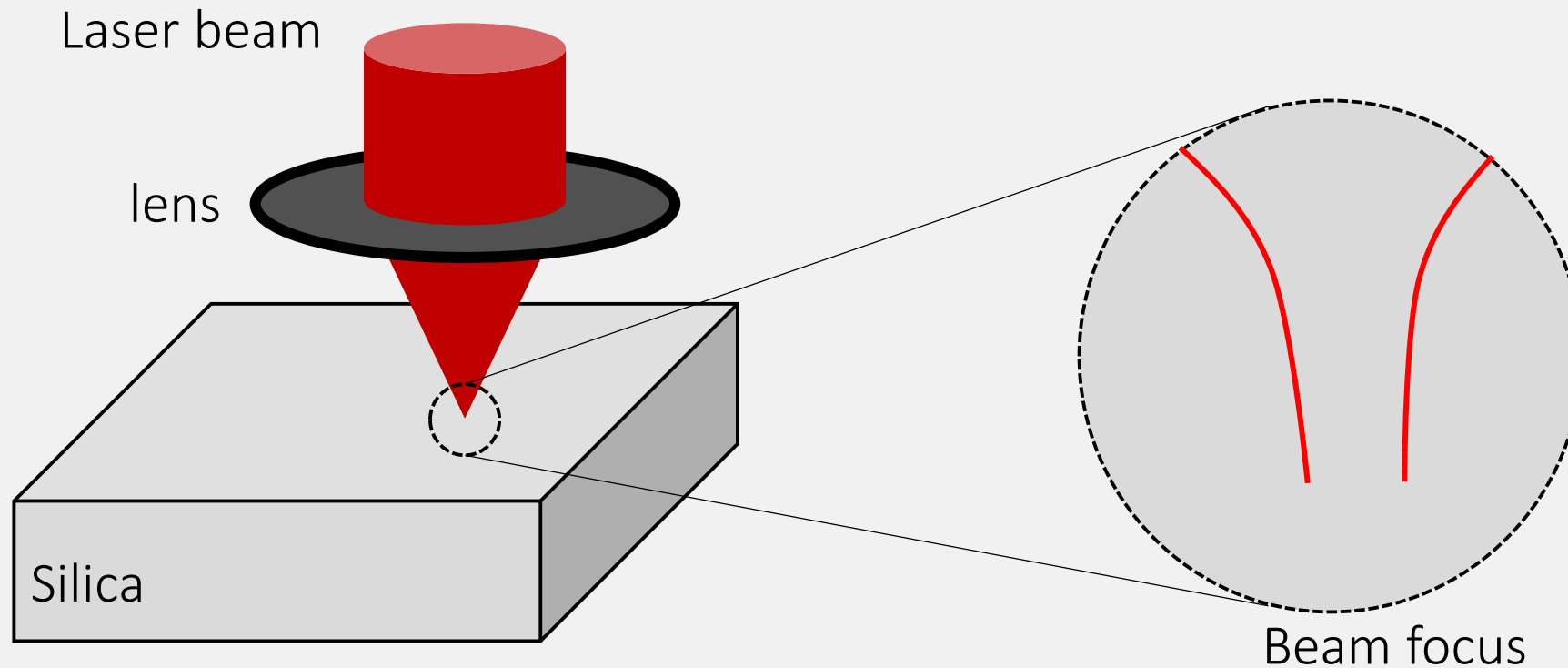
Femtosecond (10×10^{-15} s) laser
induced *small* gratings in **quartz glass**.



Three-dimensional optical storage inside transparent materials. E. N. Glezer, M. Milosavljevic, L. Huang, R. J. Finlay, T.-H. Her, J. P. Callan, and E. Mazur . Optics Letters Vol. 21 Issue 24 (1996)

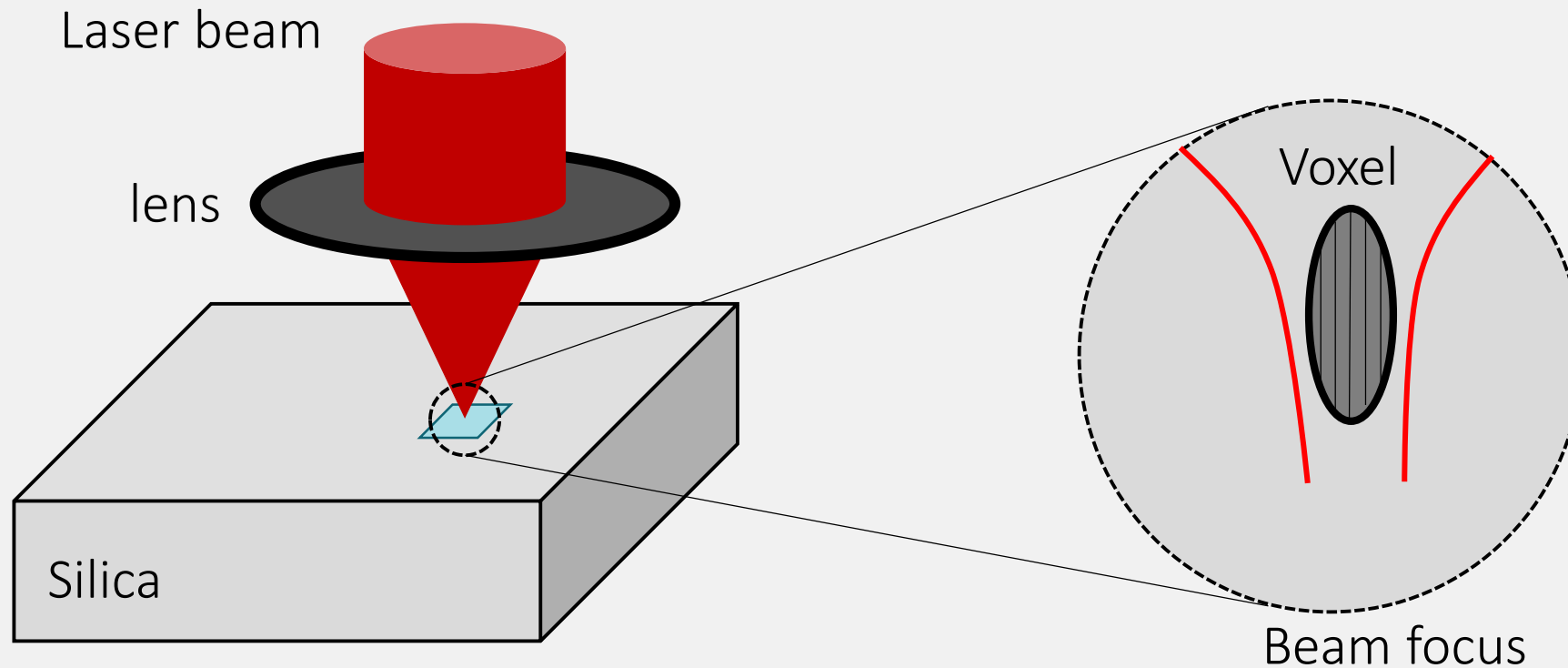
Writing in Silica glass

- Uses an ultra-short pulsed laser
- The beam is focused onto the glass



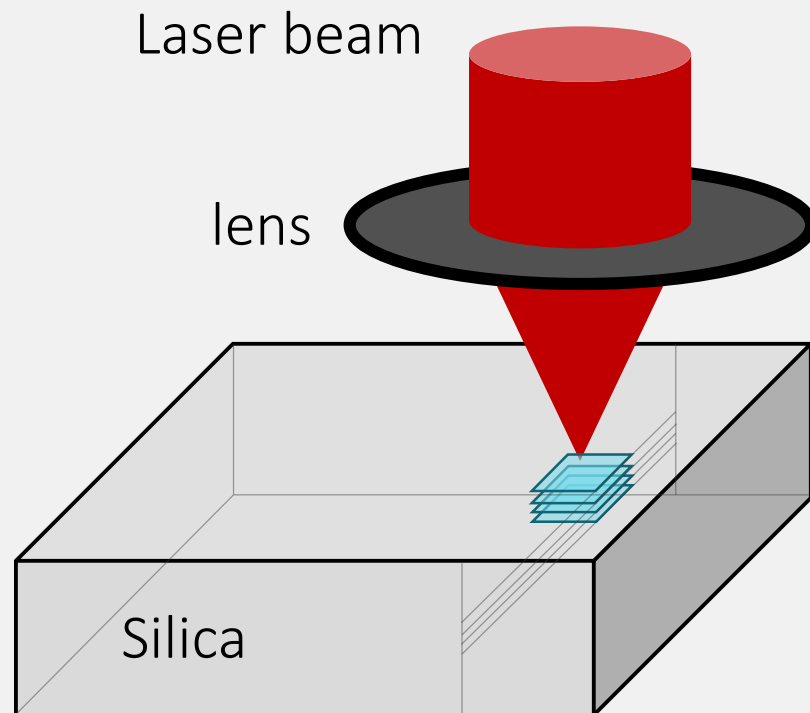
Writing in Silica glass

- The material is modified at the focus
- A voxel is formed
- Repeat, to form a sector of voxels



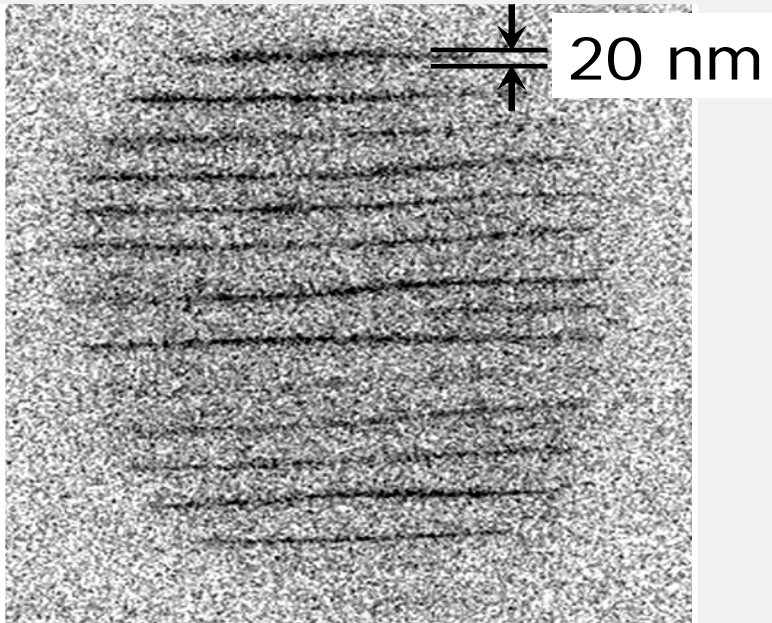
Writing in Silica glass

- Sectors are written in bulk of material, protecting the voxels



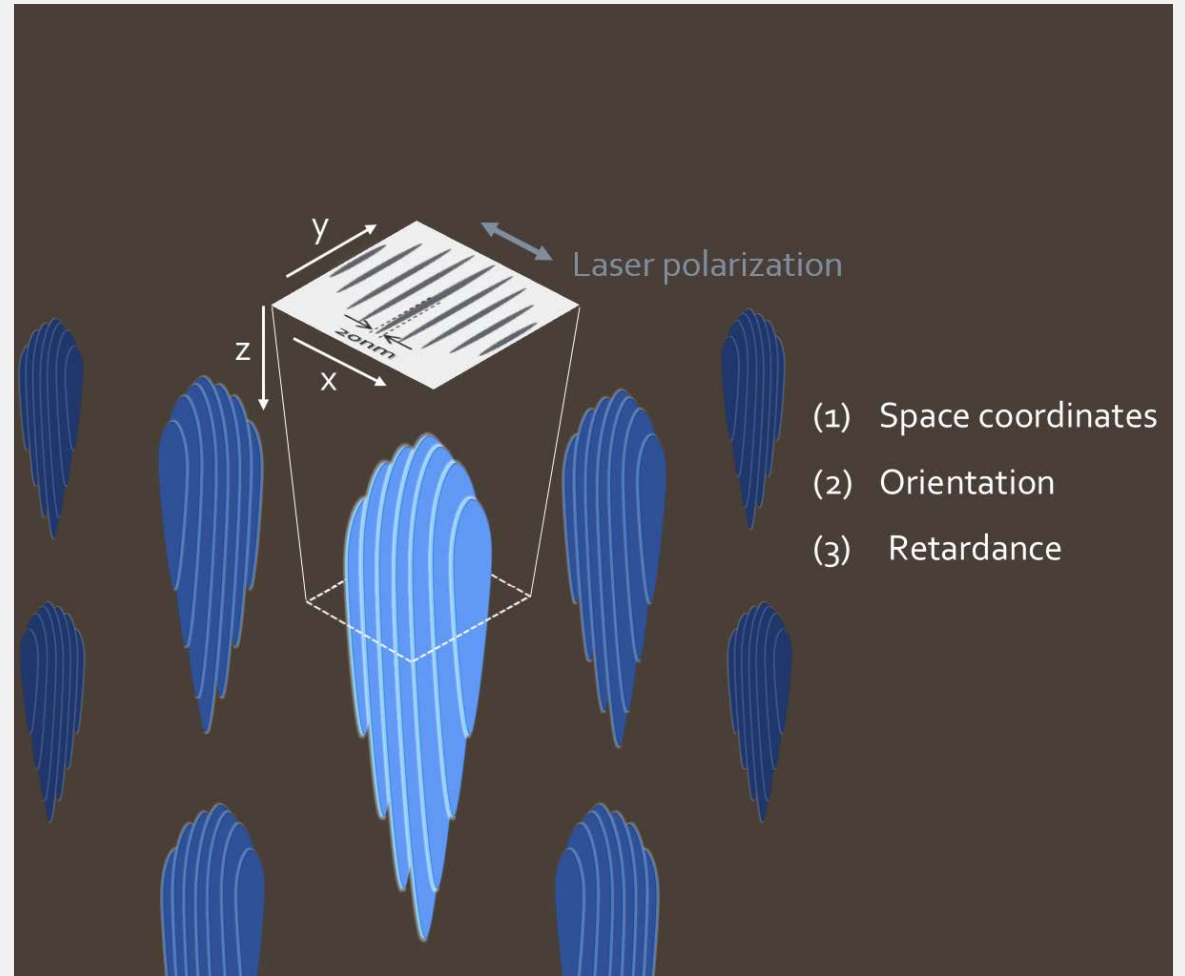
Nano-grating structure and control

- Multi-level encoding using shape and orientation of structure



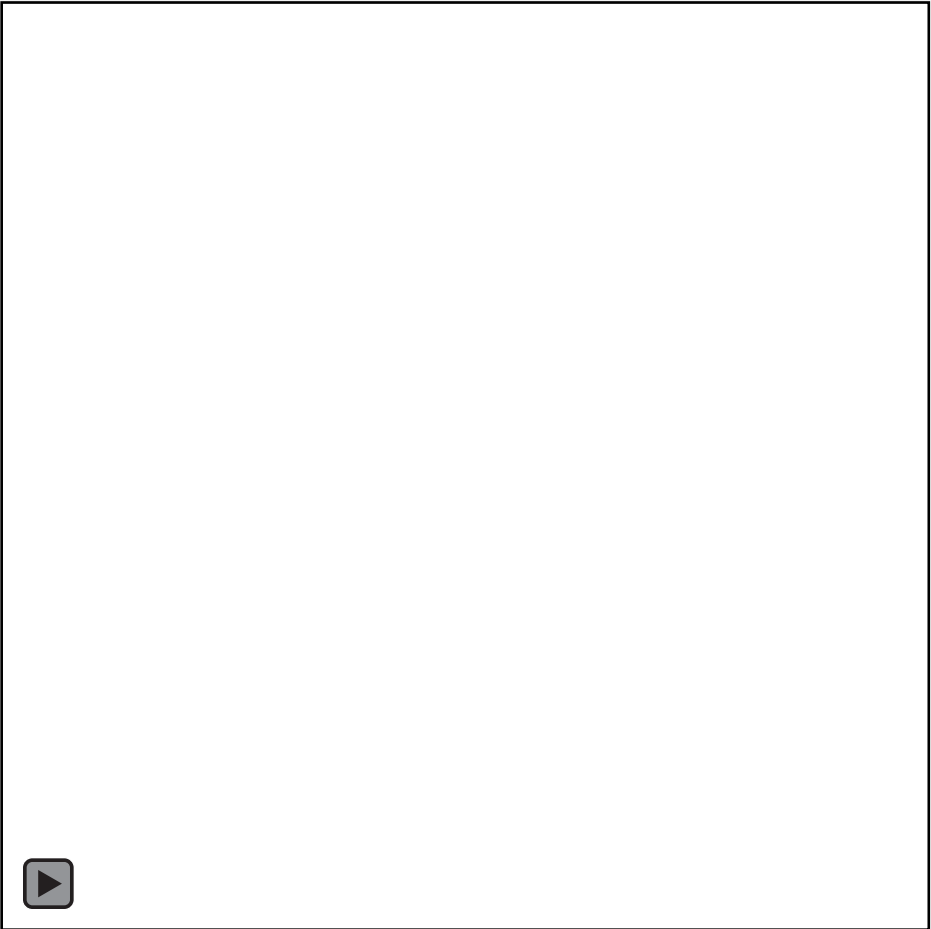
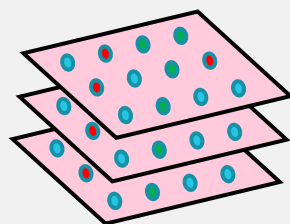
Self-Organized Nanogratings in Glass Irradiated by Ultrashort Light Pulses
Yasuhiko Shimotsuma, Peter G. Kazansky, Jiarong Qiu, and Kazuoki Hirao
Phys. Rev. Lett. **91** (2003)

100's of layers

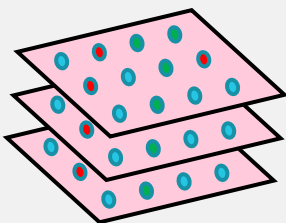


Writing voxels

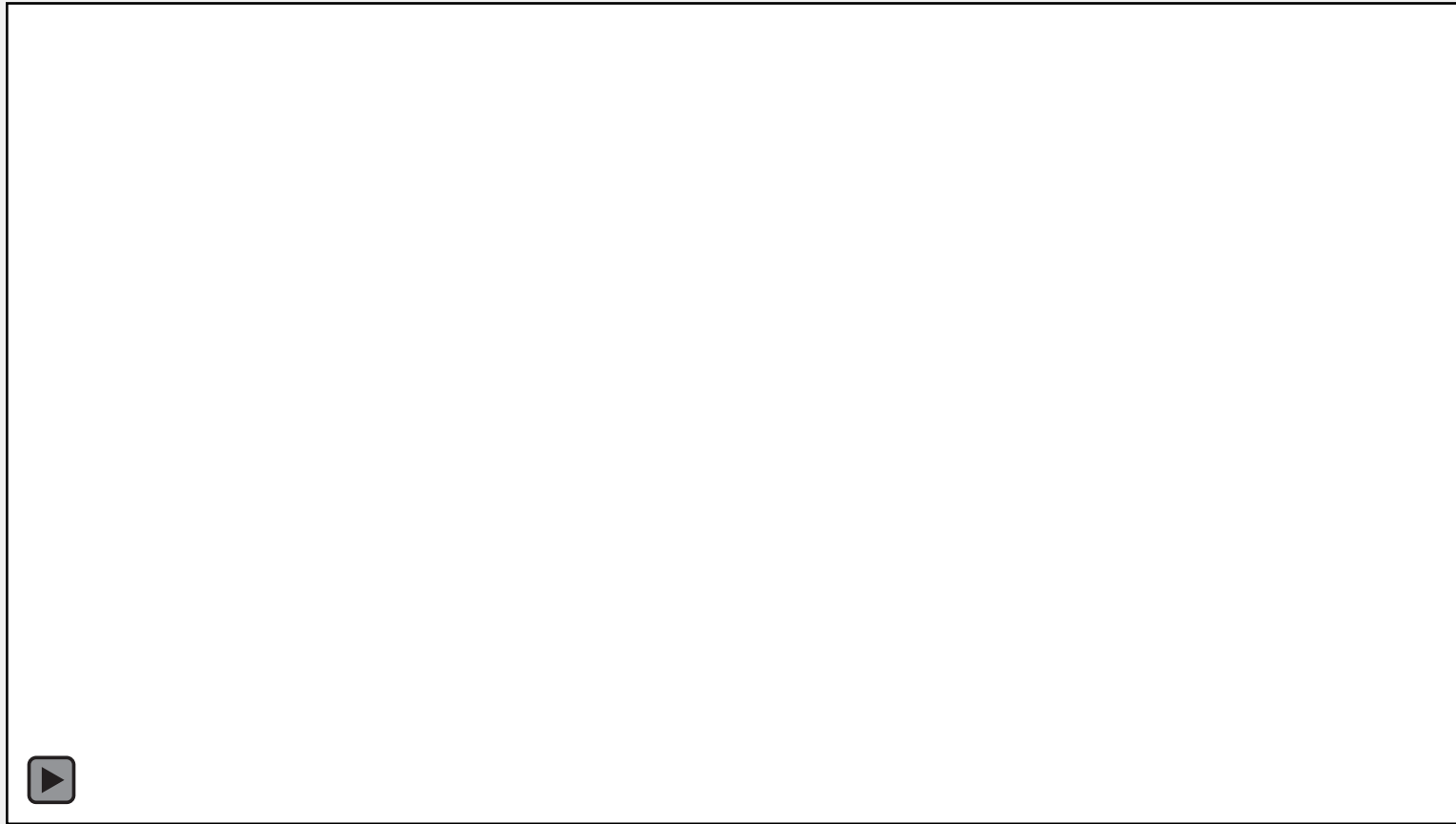
Sectors
of voxels



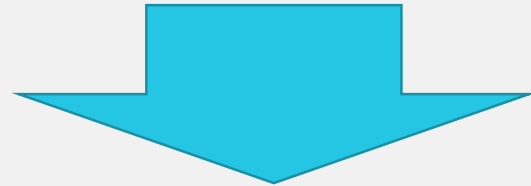
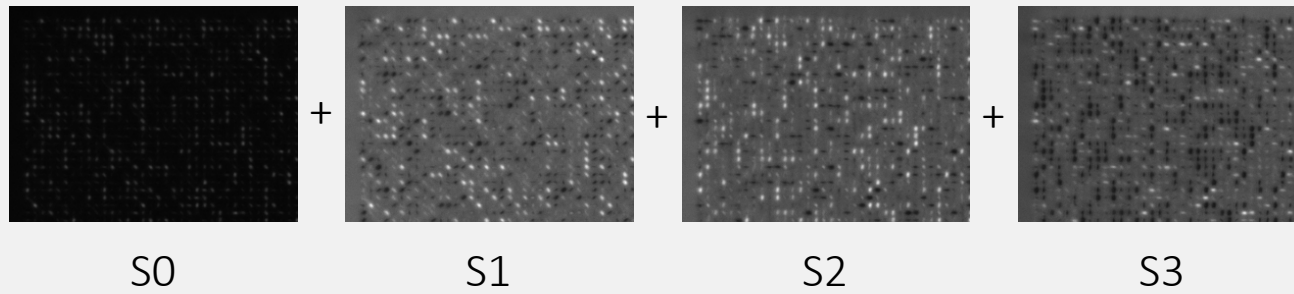
Sectors
of voxels



Reading voxels

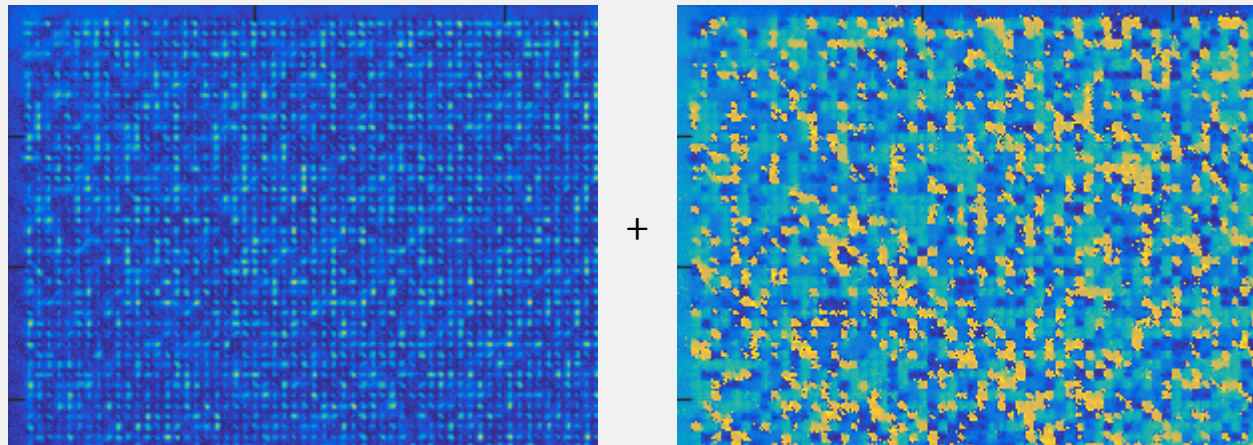


Decoding voxels

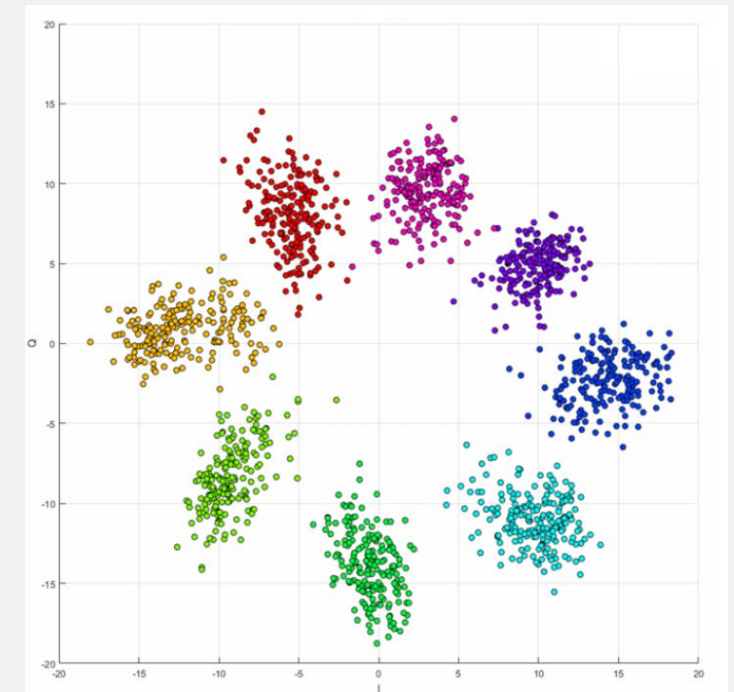


Retardance

Orientation



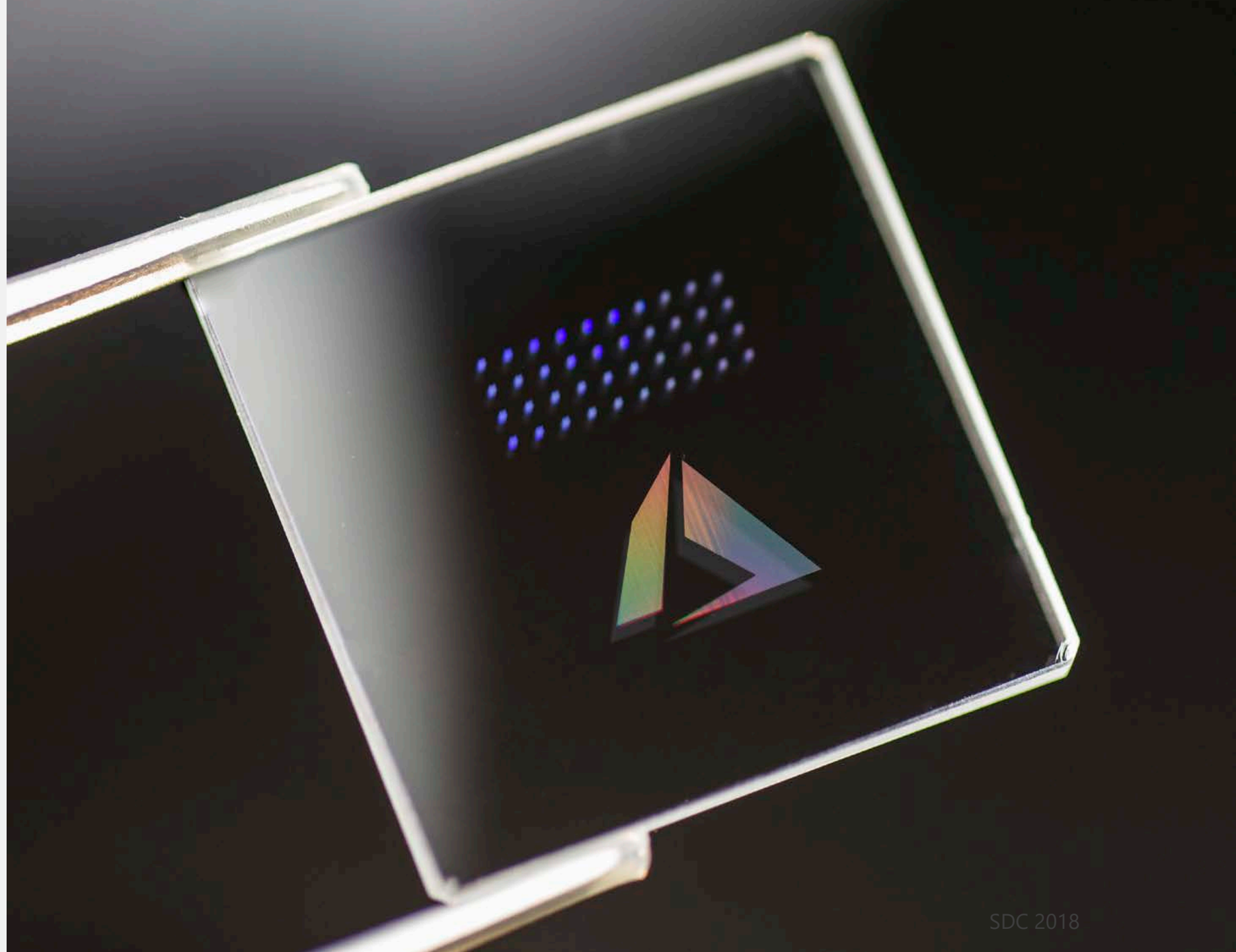
Orientation +
Retardance



Constellation diagram


Summary


- Glass is:
 - Archive-grade
 - WORM-like
 - Seekable
 - Cheap
- Excellent lifetime





Thank You

 Austin Donnelly

 austind@microsoft.com

 www.microsoft.com