Feeding the Pelican

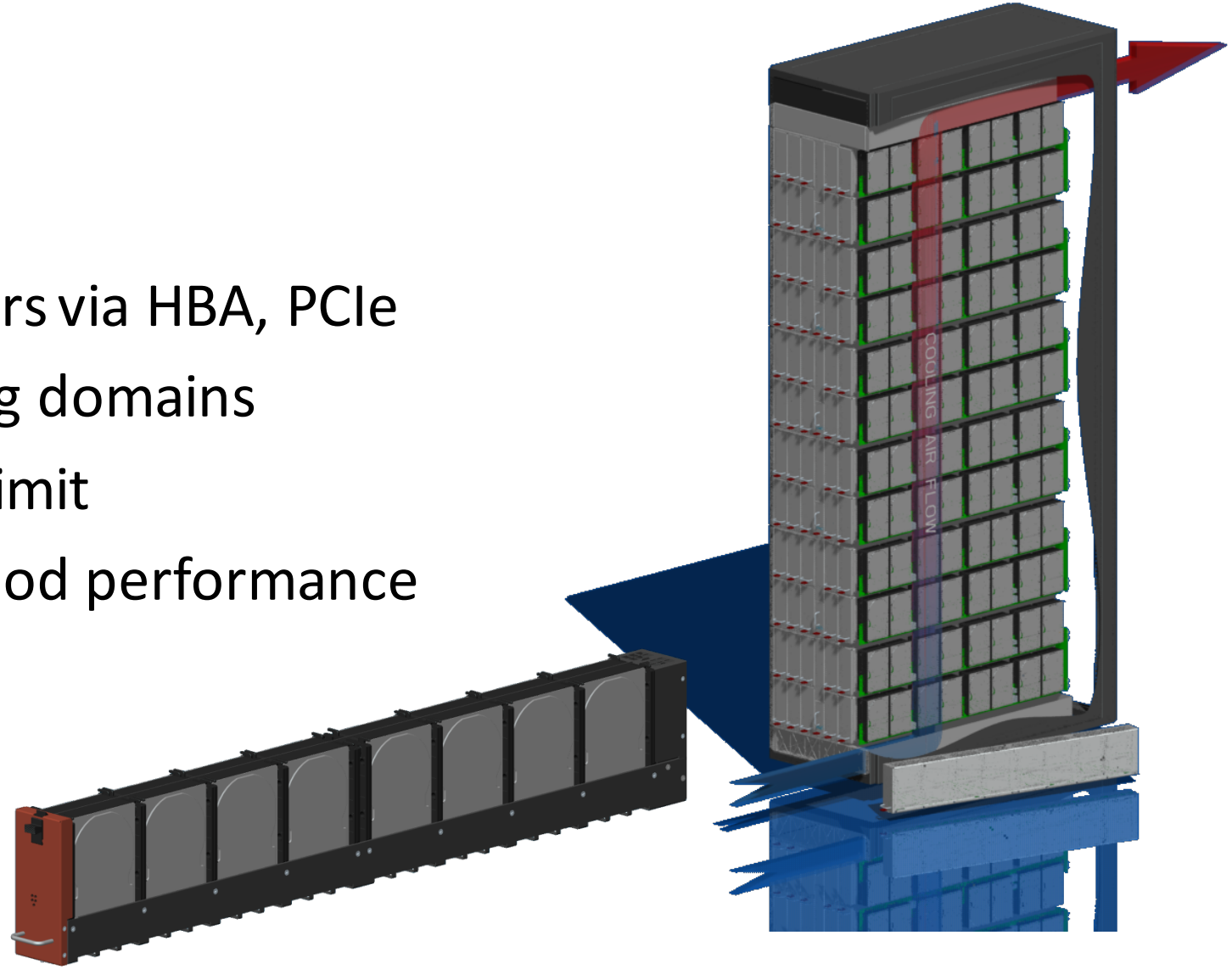
Using archival hard drives for cold storage racks

Austin Donnelly[†], Richard Black[†], Dave Harper[†], Ant Rowstron[†], Aaron Ogus[‡]

[†]Microsoft Research, [‡]Microsoft

Pelican

- 1152 Archive-grade HDDs
- Directly attached to 2 servers via HBA, PCIe
- Orthogonal power & cooling domains
- Spundown to meet power limit
- Schedule requests to get good performance



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

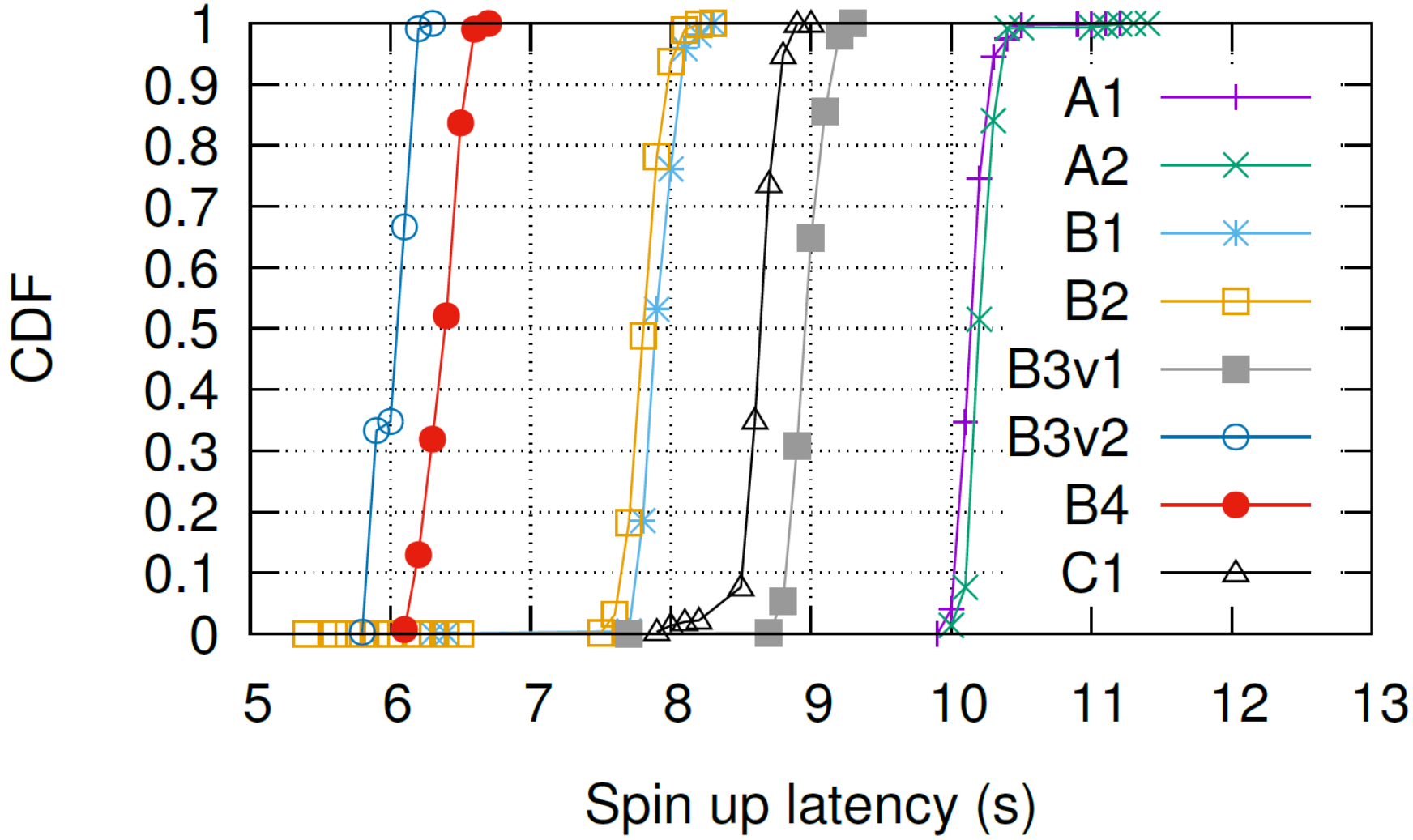
- Workload is quantified as TB/year
- Lifetime affected by:

  POH TB transferred Spindown cycles

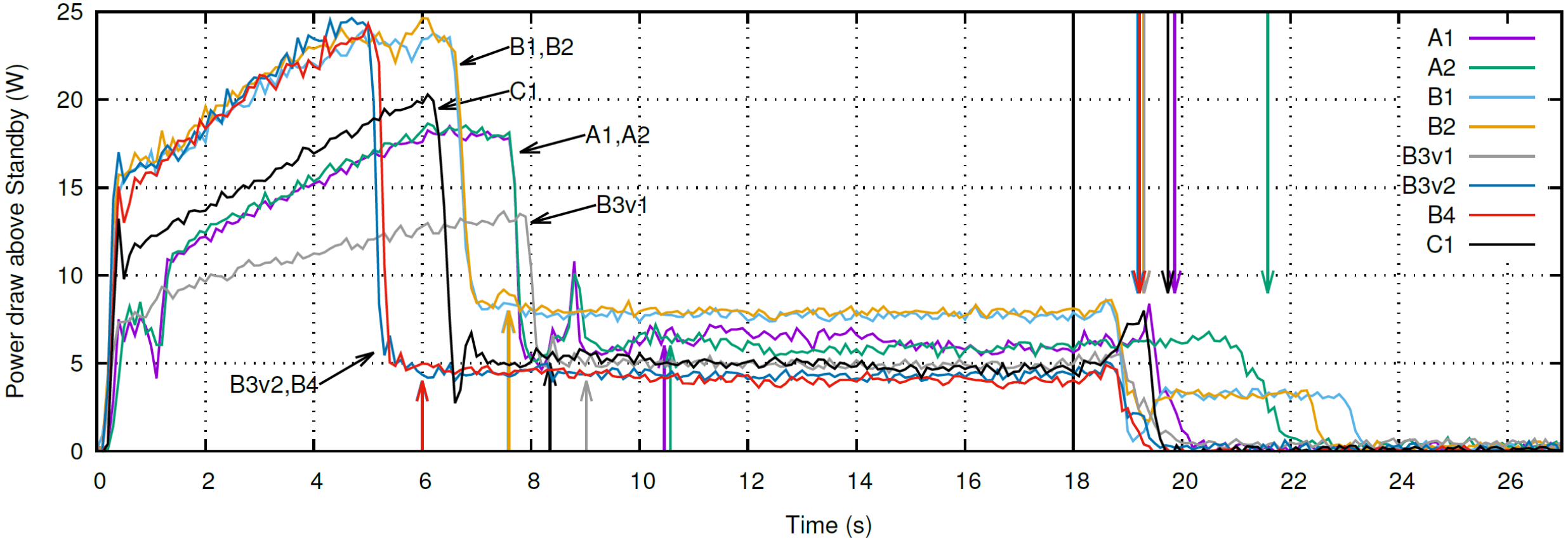
Drive line-up

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

Spinup latency

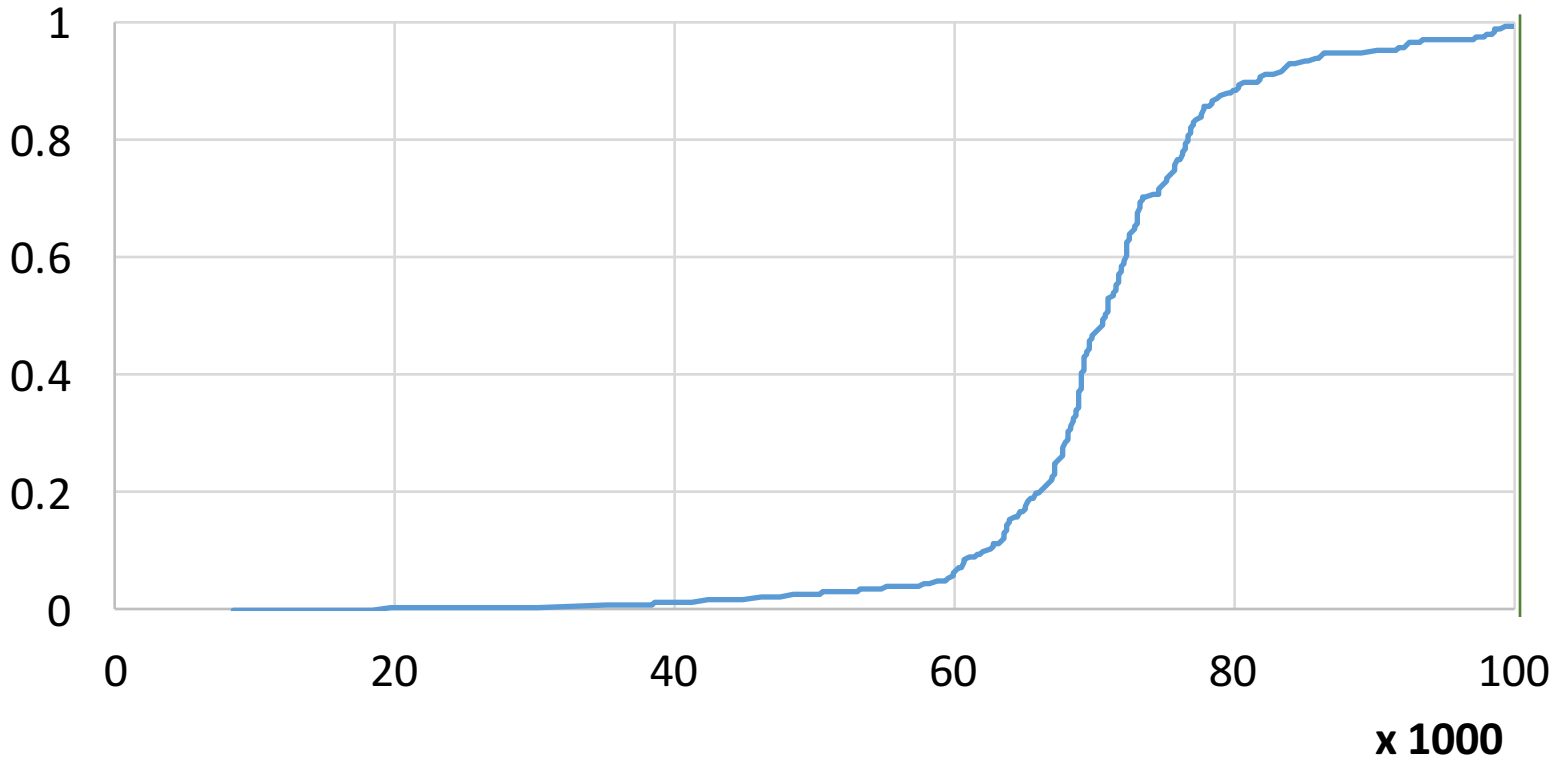


Power draw


Power draw above Standby (W)
Time (s)
B1,B2
C1
A1,A2
B3v1
B3v2,B4
A1
A2
B1
B2
B3v1
B3v2
B4
C1

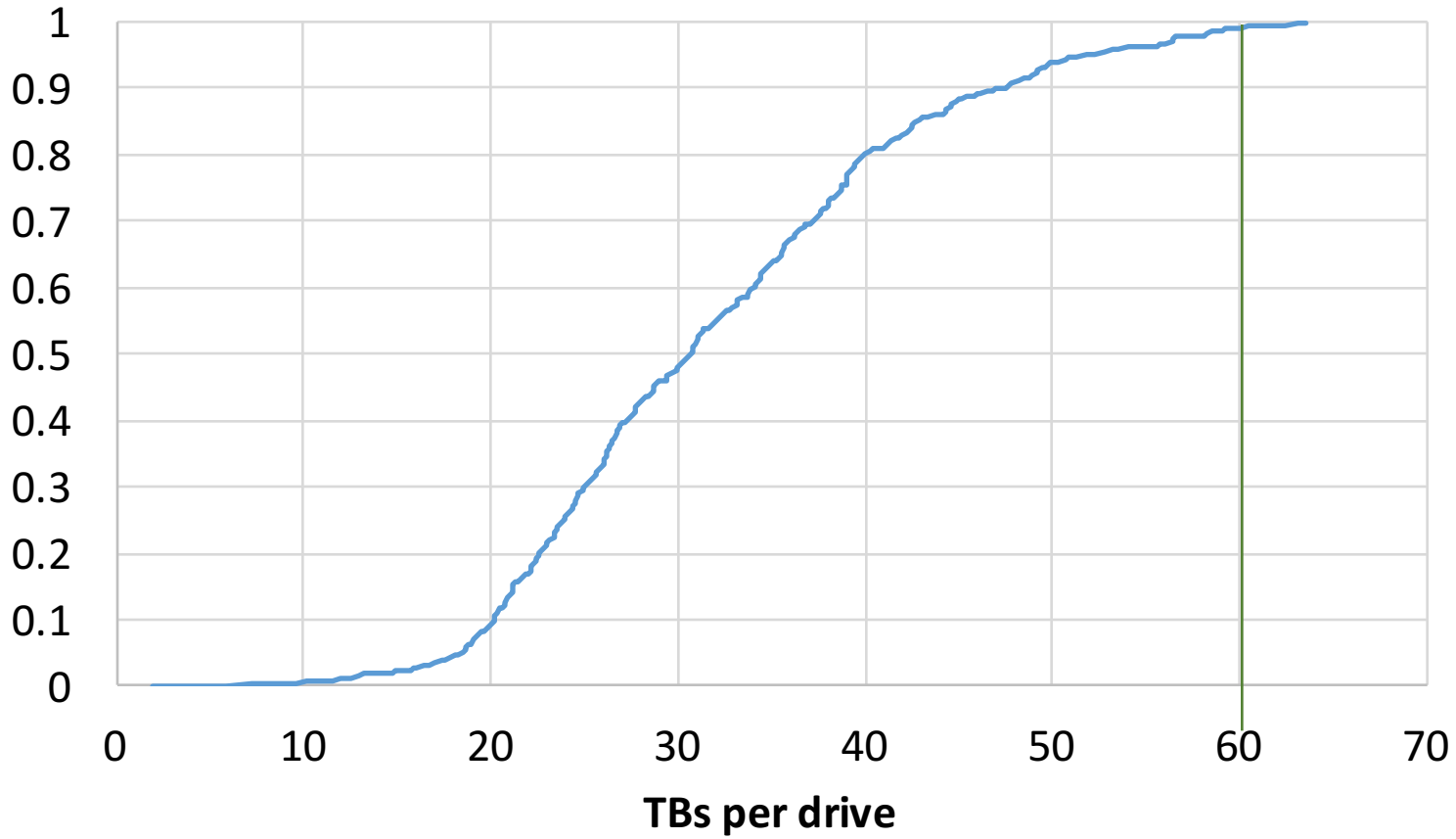
Is it ok to do all these spinups?

datasheet spec: 100K per year.



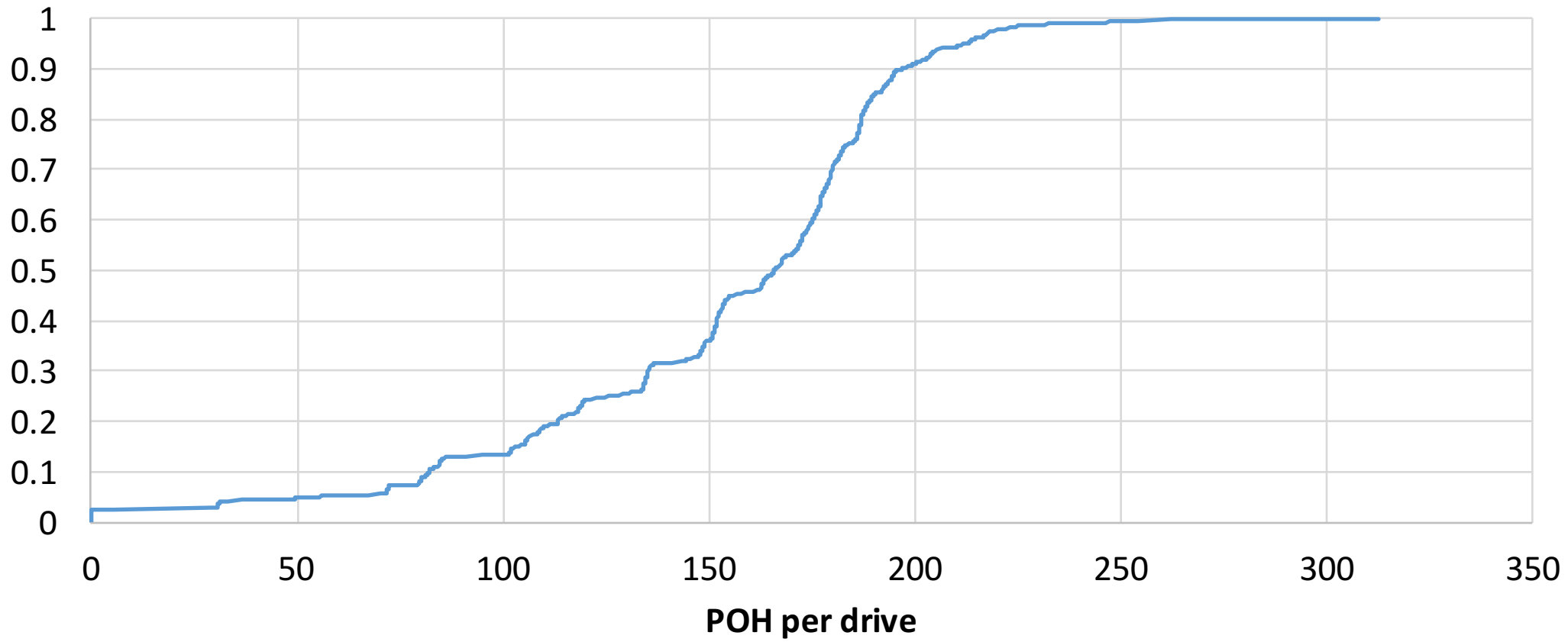
TBs transferred

datasheet spec: 60TB/year



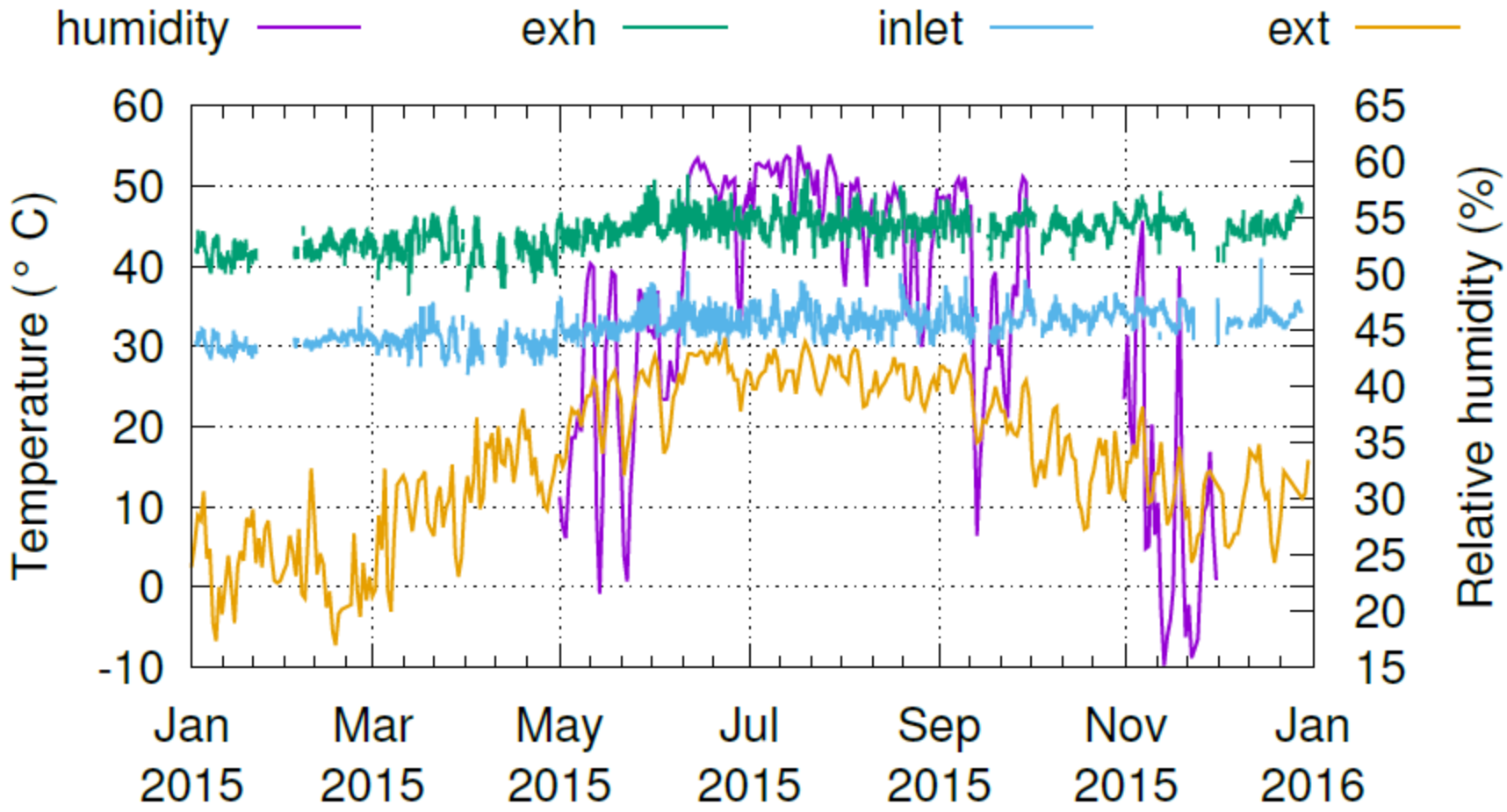
Power On Hours

datasheet spec: 3120 POH/year (about 1/3rd of a year)

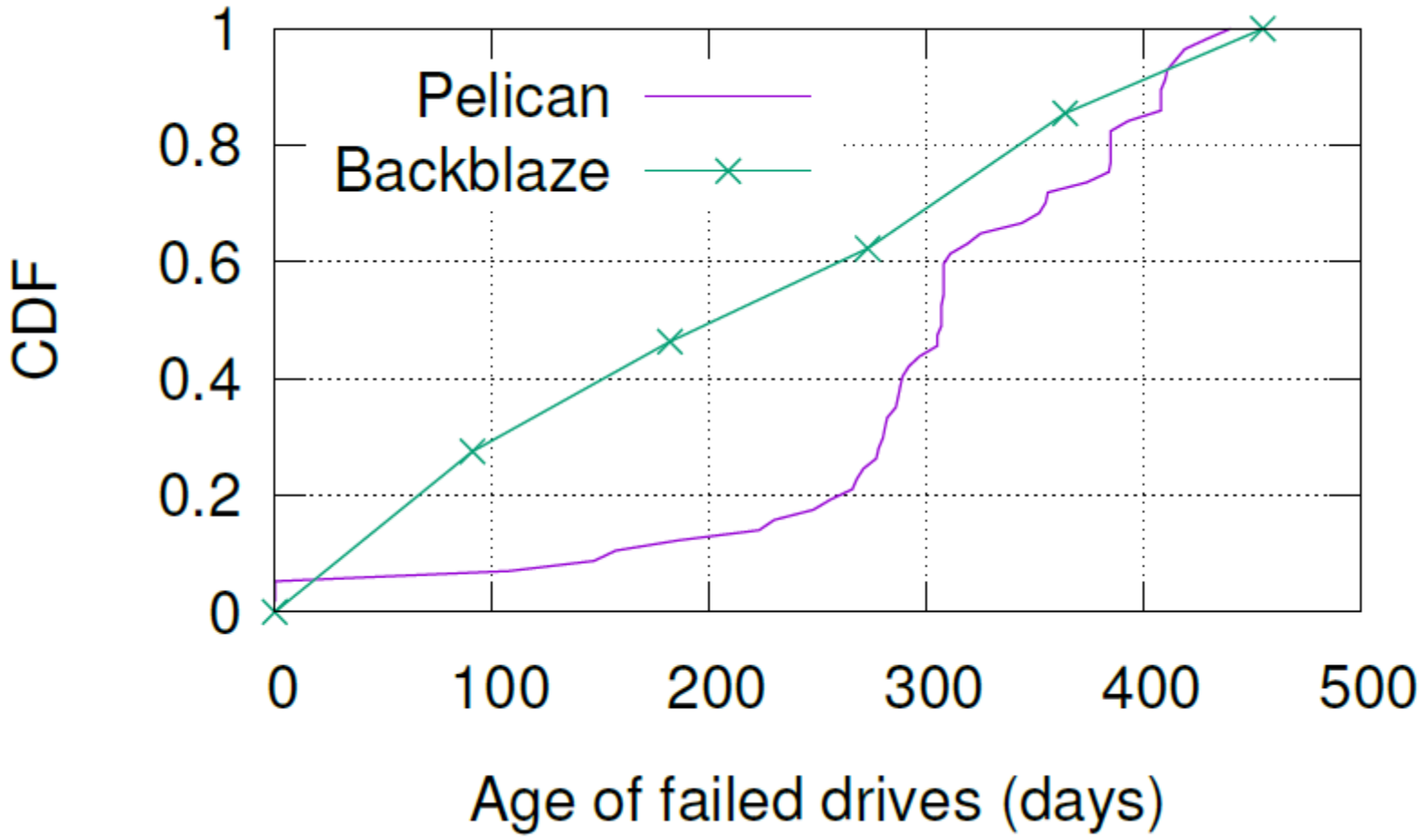


Temperature

Datacenter environment



Failure rate



AFR: 3.96%

Conclusion

- Archive drives are effective, provided workload is managed
- Spindowns do **not** seem to affect drive reliability
- Temperature and humidity are bigger factors

- Future:
  - Workload management not just for SSDs
  - Worth understanding drive performance at firmware level
  - Power control is critical
  - Background operations control is needed too (particular for SMR)