

From Leader to Fast Follower through the Barren Cloudy HPL Valley

02/2025

Gary Grider

Los Alamos National Laboratory

LA-UR-24-28694

Pre-cloud era

Of course, there was the pre cluster era and the pre integrated circuit era and the pre-transistor era but this examination is after those early eras

- HPC/Simulation/Analytics was the big fish leading technologies
 - Process parallelism with some local data parallelism was common
 - HPL was somewhat useful in guiding industry but only because clock rates were king
- Administrative computing was all on prem and either nice home-grown software or big-name software packages were the norm

The Barren Cloudy HPL Valley

- Cloud was the big fish leading technologies
- Cloud technologies were fine for administrative computing and off prem capex->opex and subscription software as a service became the norm
- HPC/Simulation/Analytics entered an era of slipping into the most inefficient state one could imagine
 - Clock rate increases ended, many core and data parallel acceleration boomed, highest HPL while efficiencies record lows efficiencies for all but embarrassingly parallel problems dense problems)
 - Exascale - the pinnacle of inefficiency with the least balanced systems setting the records
 - “I have harped on the imbalance of the machines,” Dongarra said.
 - Cloud tech drove in an opposite direction for weak scaled low-density data centers
 - HPC was not the big fish so really had little effect on the market
- HPC had no market to fast follow and was so small it couldn't shape the market
- HPC integrators start to drop like flies because cloud doesn't use their wares

The Cloud+AI era – HPC becomes a fast follower

- Cloud plus AI makes HPC a tiny niche and neither cloud or AI use an integrator
- AI factories
 - push high power, dense, network rich solutions, use their own software except parallel FS's
 - utilize dense memory access accelerators at large volumes
 - need high memory bandwidth and some sparse access for inference
- HPC can now follow AI – if HPC needs can be mapped into AI directions
 - Sparse memory access is needed by some inference and can be pushed into industry helping HPC map all its apps to denser access forms making following AI both feasible and efficient
- Few integrators – Cloud/AI factories use their own proprietary software. HPC community must provide its own broad community integration tools.
 - System mgmt, open/community helps both hpc and non factory cloud/ai sites
 - Resource mgmt, perhaps an area that HPC community could have their own broad solution
 - HPC can leverage ai factory racks/data center solutions/form factors as they want the density
 - HPC sites must leverage containers to push all snowflake costs to the edge (apps)
 - Data mgmt will need to be invested in given the importance of data to all this new world

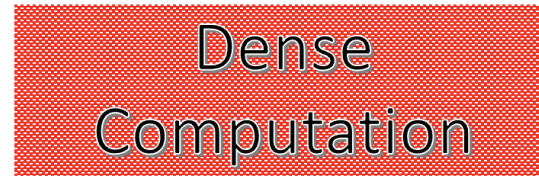
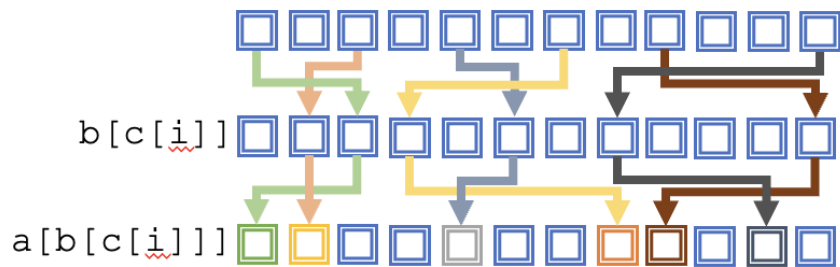
The Cloud+AI era – HPC becomes a fast follower (cont)

- This fast following a market that moves faster than HPC and has godlike funding has its downsides, staffing/YRE
 - Staff can double their compensation, and don't have to even leave their homes
 - Attracting staff might get easier because of the AI pull but keeping the best may be impossible, further need for broad community solutions that service more than just HPC and AI – even service some of enterprise, etc.
 - Replacing the top tier performers may just not be feasible
- Parallel storage systems are now used by way more than just hpc, PFS's are broadly needed but lustre is old tech and wasn't built for broad workloads
- Proprietary solutions emerging but they are proprietary and lock in oriented
 - PNFS looks like the only solution that is within a reasonable distance from truly open src and open standard / community but needs work
 - Daos looks interesting but seems a very very long shot for broad adoption unless you just use the tech as part of a more complete solution

Some examples of activities to prepare for fast following.

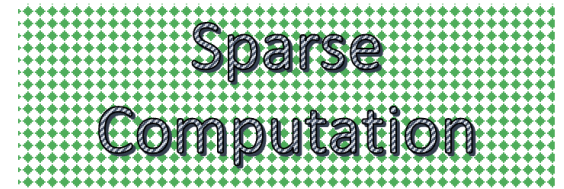
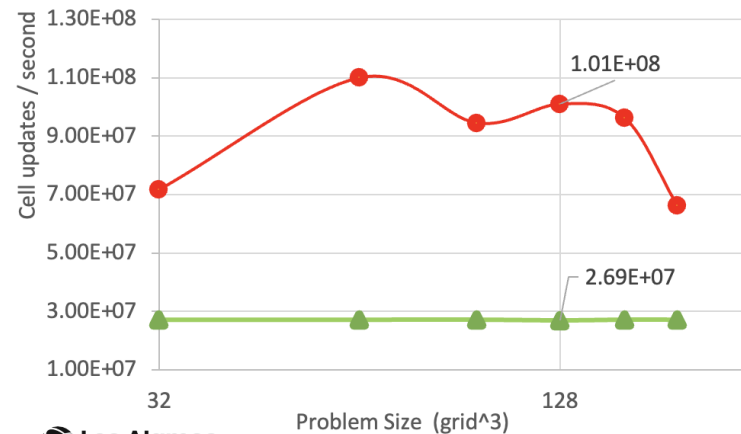
Sparse Memory Access Acceleration

- Turn our workload into AI hardware friendly , if we can find a big enough market need to help us (inference, database sparse join, etc.)
- Joint work with SK hynix



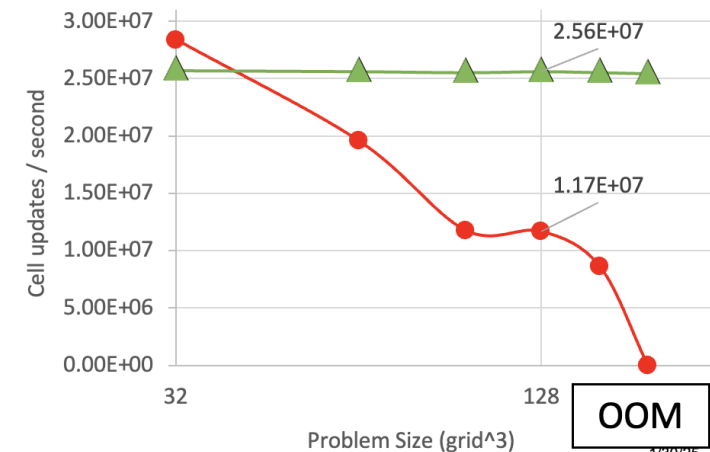
Parthenon-VIBE - AMR Hydro Proxy
32³ block size 2 AMR levels

- H100 + Grace / Grace Hopper
- ▲ Grace Superchip (2X72 cores)

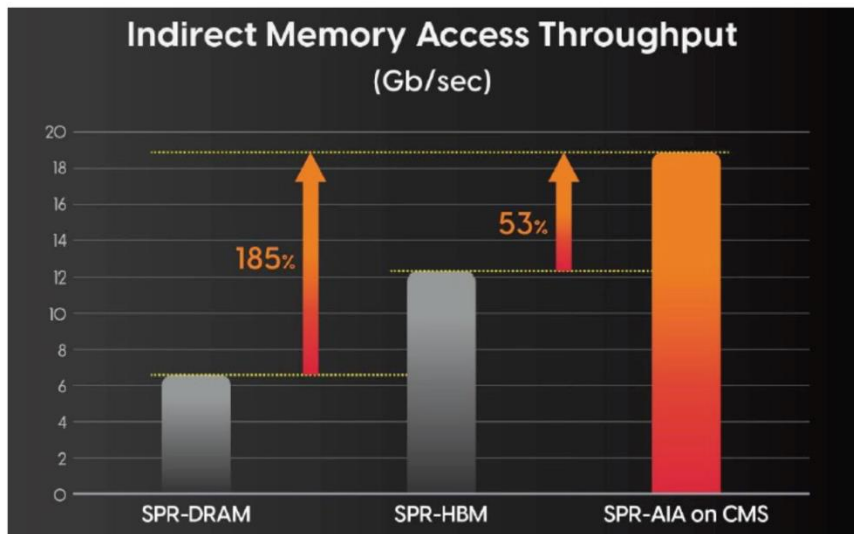


Parthenon-VIBE - AMR Hydro Proxy
16³ block size 3 AMR levels

- H100 + Grace / Grace Hopper
- ▲ Grace Superchip (2X72 cores)



OOM

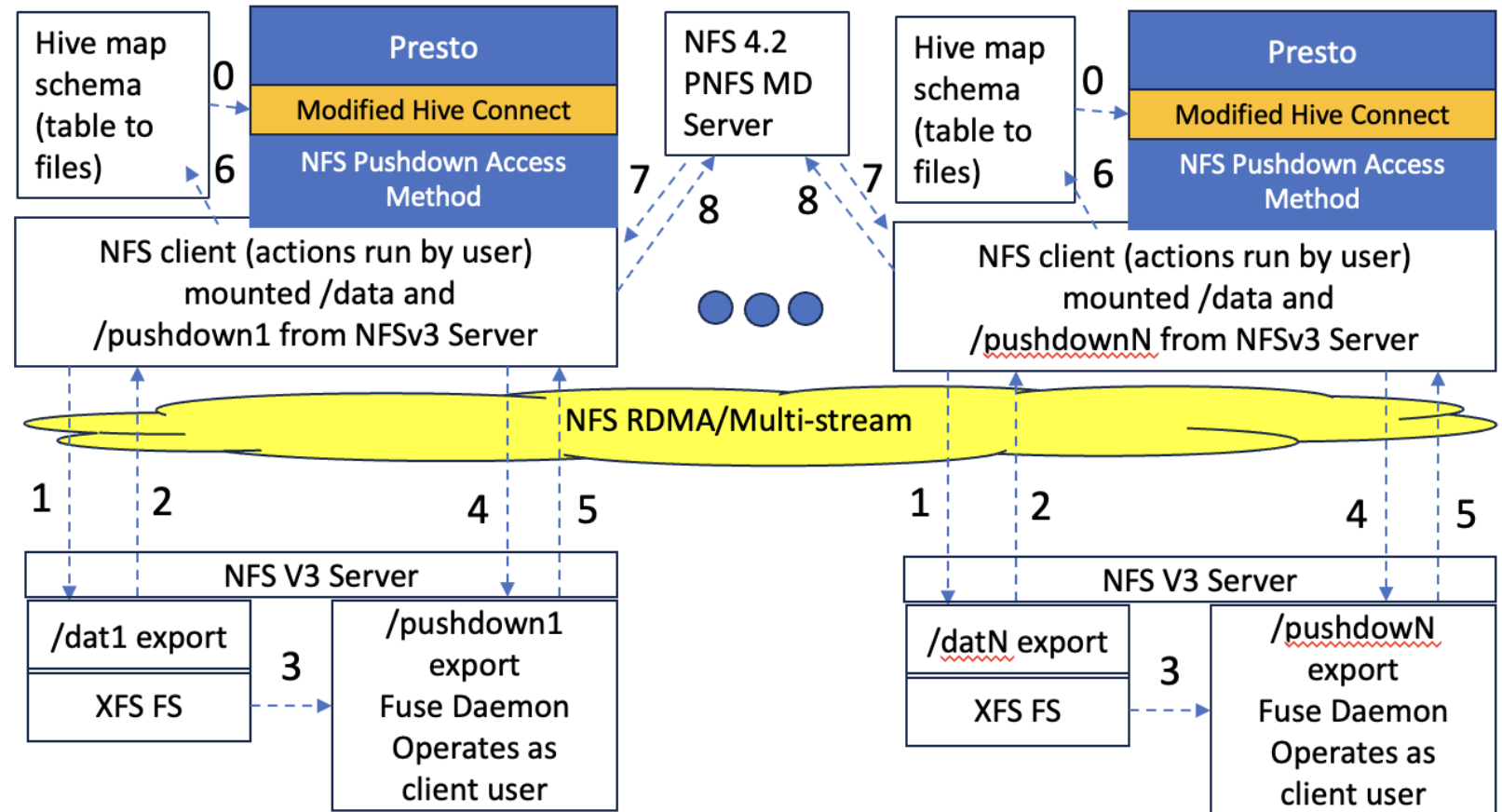


Pushdown Analytics – minimize data movement

- Leverage another large community – Apache Analytics
- Joint work with SK hynix, Hammerspace, Eideticom, Airmettle, Neuroblade, Voltron, NimbusData

Concept for upcoming Demo at ISC June 2025

Asteroid simulation
Paraview
VTK
Presto
Substrait
Arrow
PNFS
Virtual Parquet



Finding ways of bridging our world and the AI/Cloud world

- Object to file gateway
 - Object accesses have access to wealth of file/other storage legacy
 - Versity open src object gateway
- Need to know enhanced LLM
 - If you train on all data, you can't enforce need to know – posix permissions or other separation of concerns
 - LANL GUF1 is a way to have a “view” of a file system personalized to what a person can see.
 - Working with startup to produce LLM query that uses GUF1 to honor Posix need to know – dynamic vector database

Next generation extremely broadly used/supported PFS

- CMU/PDL PNFS test and eval center spinning up
- HPDA demonstration at ISC25

- Hammerspace leading the way
- At least 5 other companies engaging

PNFS/PNFS+

Open linux distro client

Open linux distro smart client

Multi net proto path client balance
failover

Open linux distro metadata server

N writers to N files

Open linux distro data server

N writers to 1 file

Being worked
now

Client Parity N writes to 1 file

Multi-directory scaling

Being worked
now

Client Parity N writes to N files

Single-directory scaling

Open (non proprietary) parallel
pushdown

Native Random File read storm

**Forming an open using
and tech providing
consortium to enable
the PNFS ecosystem**

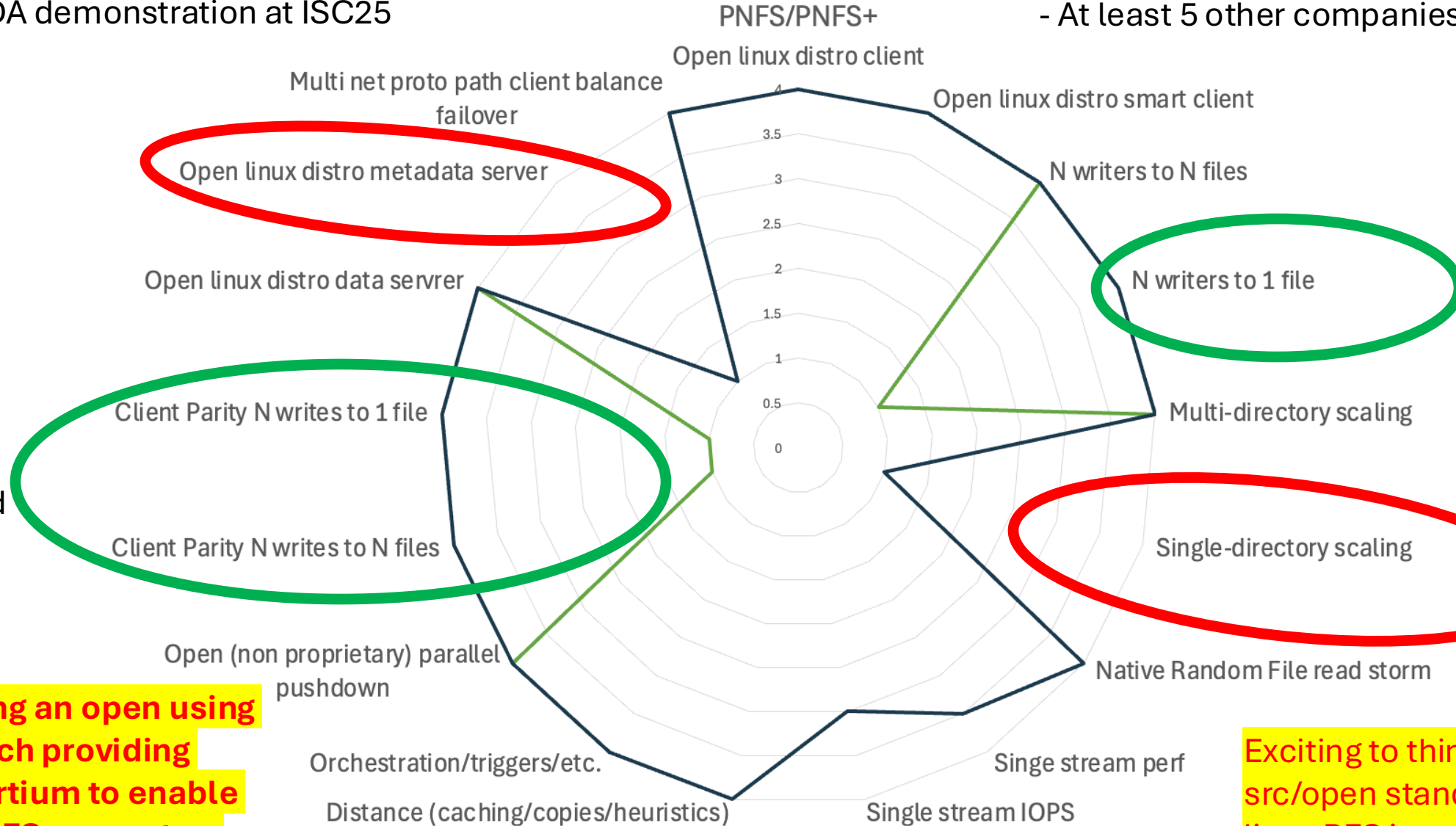
Orchestration/triggers/etc.

Singe stream perf

**Exciting to think that an open
src/open standards/ships with
linux PFS is not that far off!**

Distance (caching/copies/heuristics)

Single stream IOPS



Promoting Collaborative Communities

- Efficient Mission Centric Computing Consortium
- A collection of using orgs and tech providers exploring collaborations to build solutions

