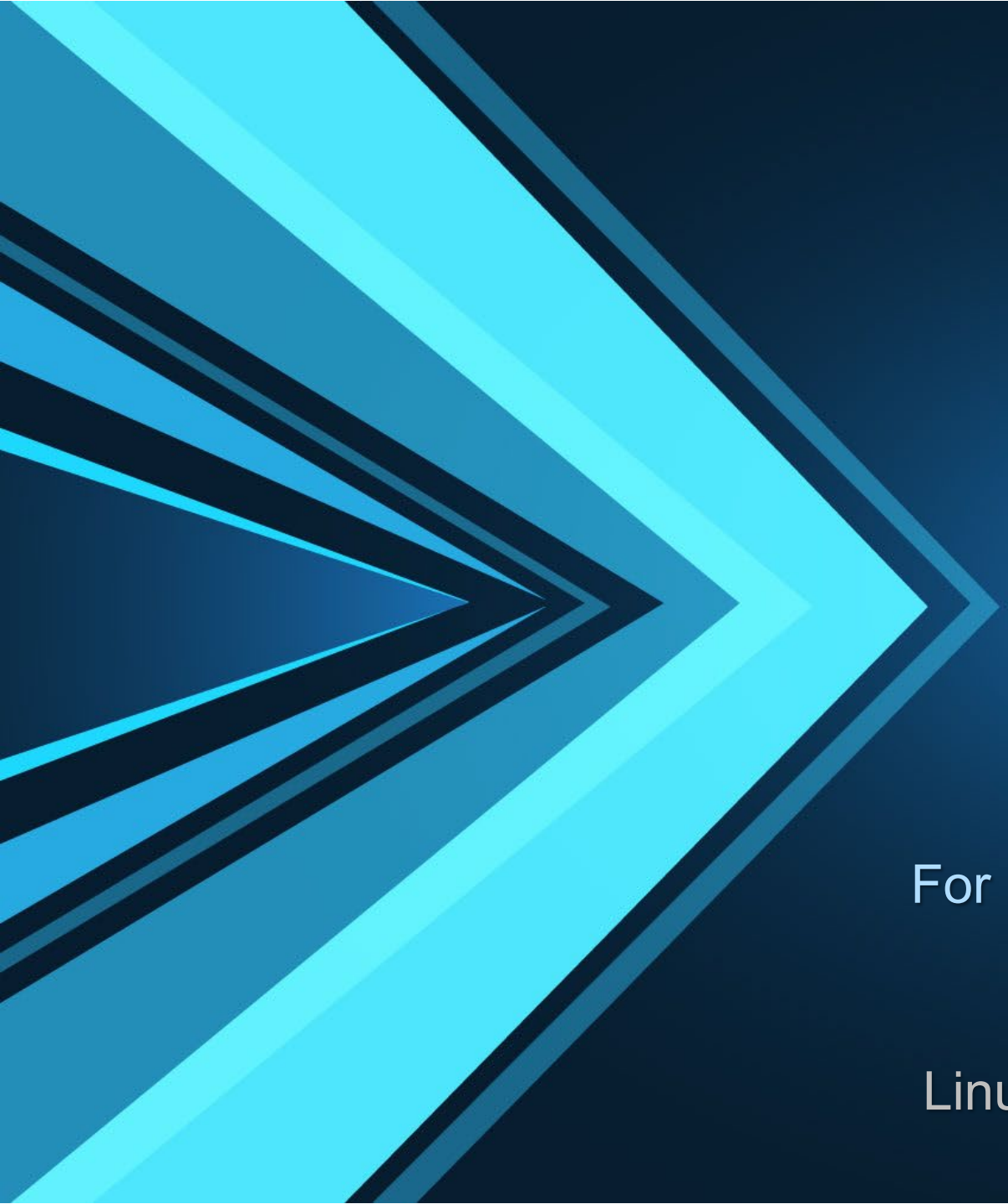


SNIA DEVELOPER CONFERENCE

SDC 24

BY Developers FOR Developers

September 16-18, 2024
Santa Clara, CA

Advancements in PNFS / NFS v4.2

For High-Performance and Distributed Storage

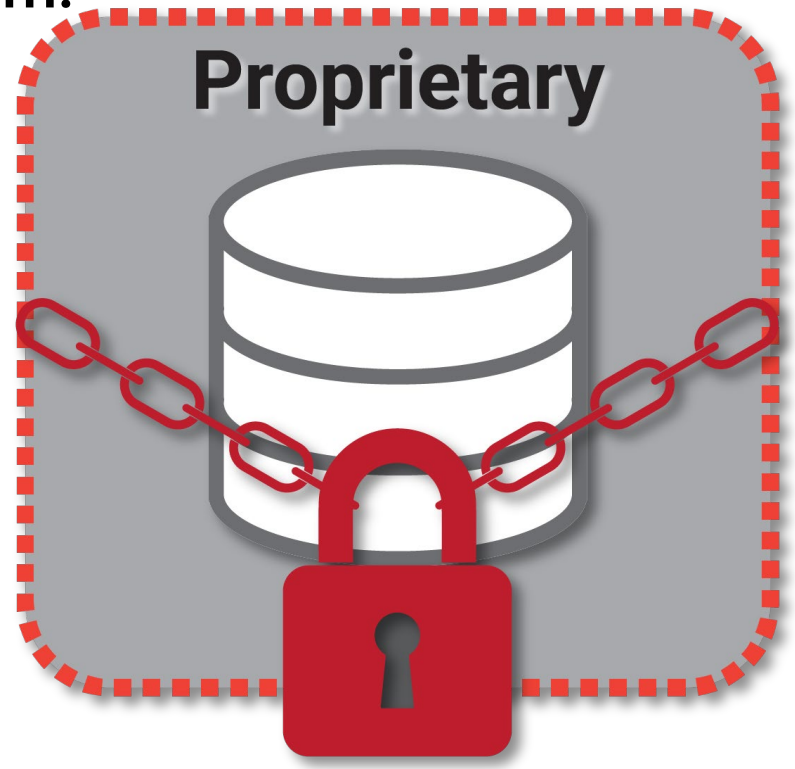
Trond Myklebust

Linux Kernel Maintainer / CTO Hammerspace

Linux Dominates in HPC and Web

AI is Driving Enterprise Adoption of HPC and Web Infrastructure Architectures

From:



To:

PNFS v4.2 at Scale: Meta's AI Research Supercluster

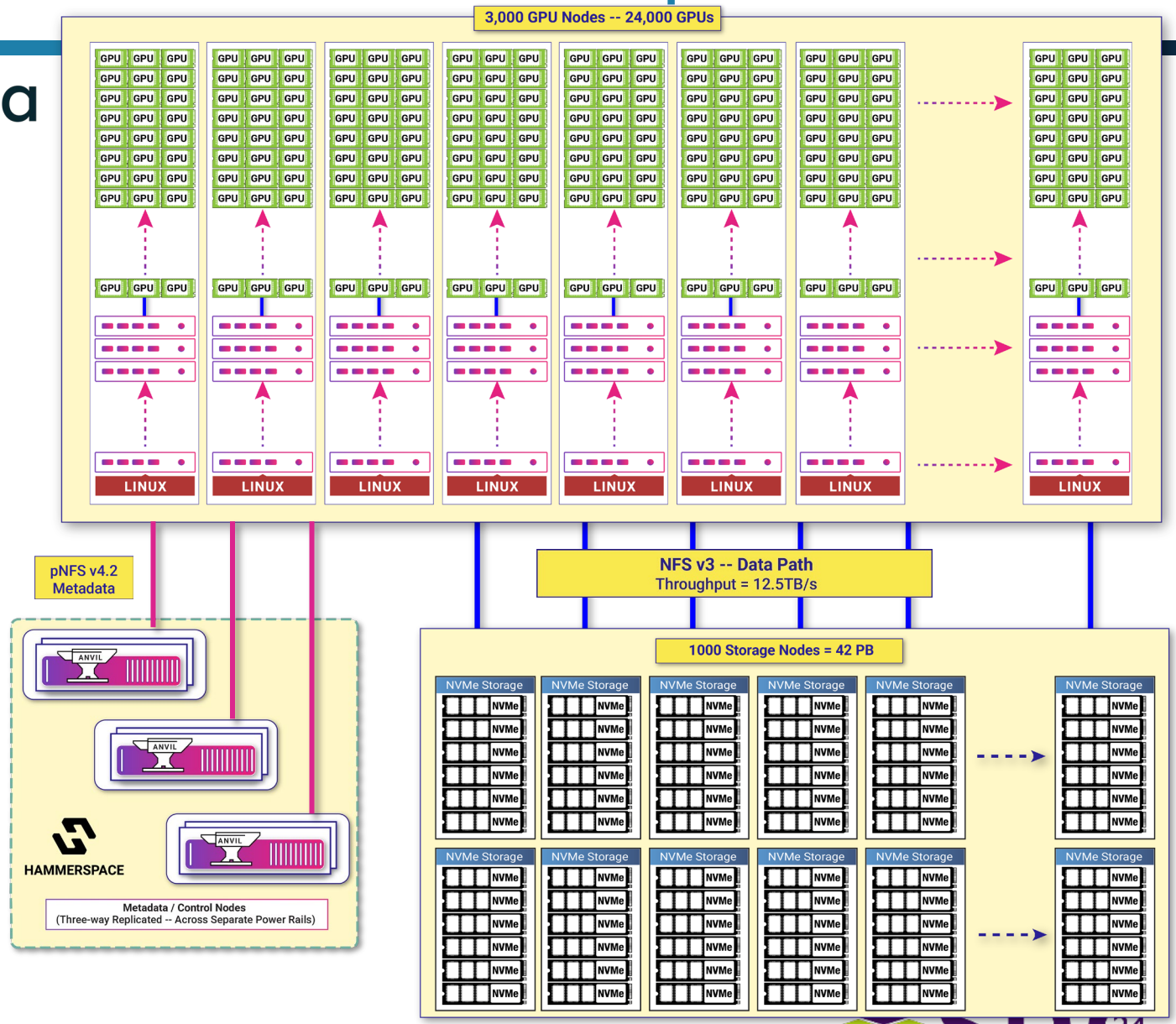
Meta

About the Customer

- Meta's AI Research Super Cluster
- Powering Llama 2 & 3 LLMs
- Massive performance and scale demands
- Evaluated leading storage vendors

Hammerspace Solution based on pNFS v4.2

- Triple redundancy on metadata nodes
- 42PB across existing **1,000+ node storage cluster**
- Feeding 24,000 GPUs, soon to be 350,000, then 1M
- Aggregate performance of **12.5TB/sec (100Tb/sec)**
- Everything is **standards-based** and **plug-n-play**
- Customer was able to use **existing OCP storage servers**



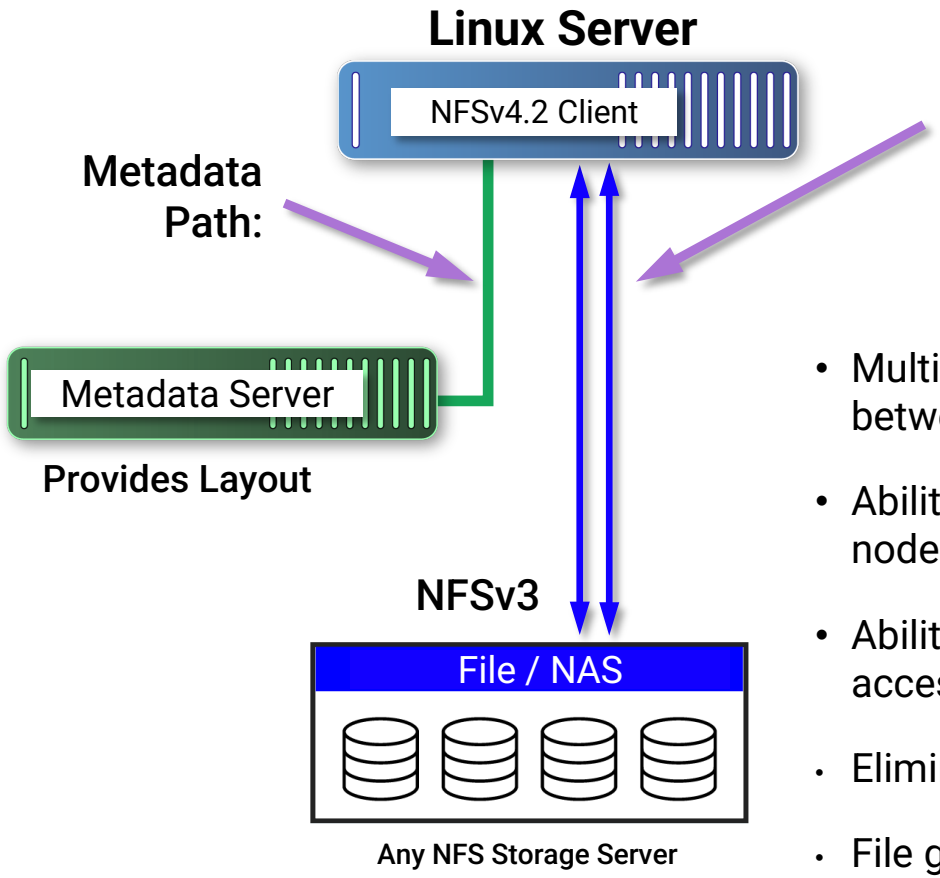
Linux Community Validating & Enhancing the Standard

- **Internet Engineering Task Force (IETF)**
  - Linux NFS enhancements are required to adhere to the IETF internet protocol standards, and often help to drive their development.
  - A multi-step process for ensuring new feature proposals get wide-range community feedback, development, review, and validation.
  - Ensures vendor-neutrality, and broad industry support.

Standards-Based Global Parallel File System Using NFS

About Parallel NFS v4.2 with Flex Files

- Parallel NFS (pNFS) introduced as optional feature of NFS in 2010, enhanced in later RFCs
- Defines a standards-based parallel file system architecture using NFS
- Architecture requires NFSv4.2 client which is part of the Linux kernel
- Provides for multiple parallel network connections between client and server



Direct data path between Linux client and storage volumes

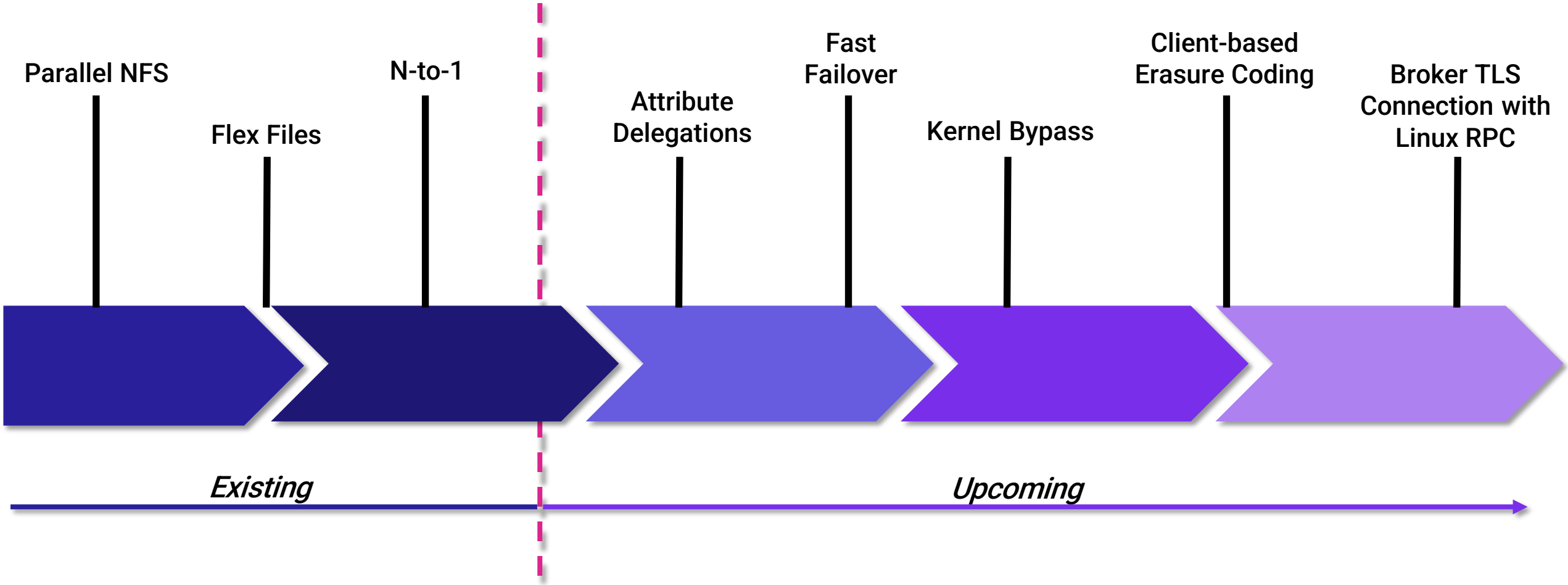
Also provides for:

- Multiple parallel network connections between client and server
- Ability to write to multiple storage nodes synchronously
- Ability to move data while it is being accessed without interruption
- Eliminates protocol chatter
- File granular access and telemetry
- TCP or RDMA Supported

NFS4.2 – Recent NFS Enhancements and Fixes

- **Elimination of excess protocol chatter using**
  - Compound operations (versus serialized)
  - Caching and delegations (including client-side timestamp generation, eliminating need to go to the server)
  - This eliminates 80% of NFSv3's GETATTR traffic
  - File open / create is one single round trip to the metadata service (vs three serial round trips for NFSv3)
  - Subsequent open and read of a file just written is ZERO round trips (vs two serial round trips on NFSv3)
- **Multiple parallel network connections between client and server and optional RDMA (nconnect)**
  - Avoids TCP stack performance limitations
- **Ability to write to multiple storage nodes synchronously (striping, mirroring)**
  - To build highly reliable, highly available systems from unreliable storage nodes
  - To distribute even a single file access across multiple back-end NFSv3 storage nodes
- **Ability to move data while it is live being accessed w/o interruption**
- **File-granular access / performance telemetry gathering and reporting**
- **Ability to serve SMB over NFS**
  - Mapping of Active Directory principals and ACLs over the NFS protocol
  - SMB extended attributes carried over the NFS protocol (future)
  - Converged file range locking (future)

Linux Advancements for High-Performance Storage



Thank You

Please take a moment to rate this session.

Your feedback is important to us.

SDC 24