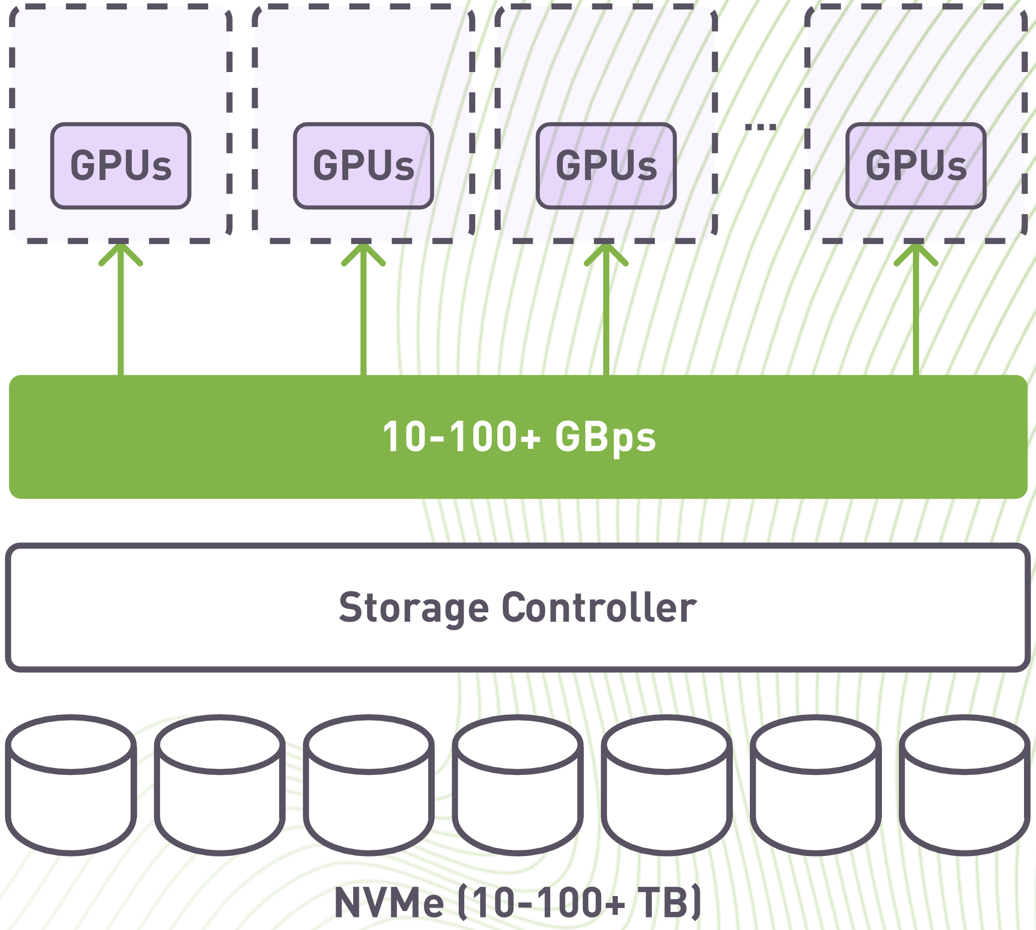
XINNOR

Distributed erasure coding for NVMe SSDs in virtualized cloud infrastructure

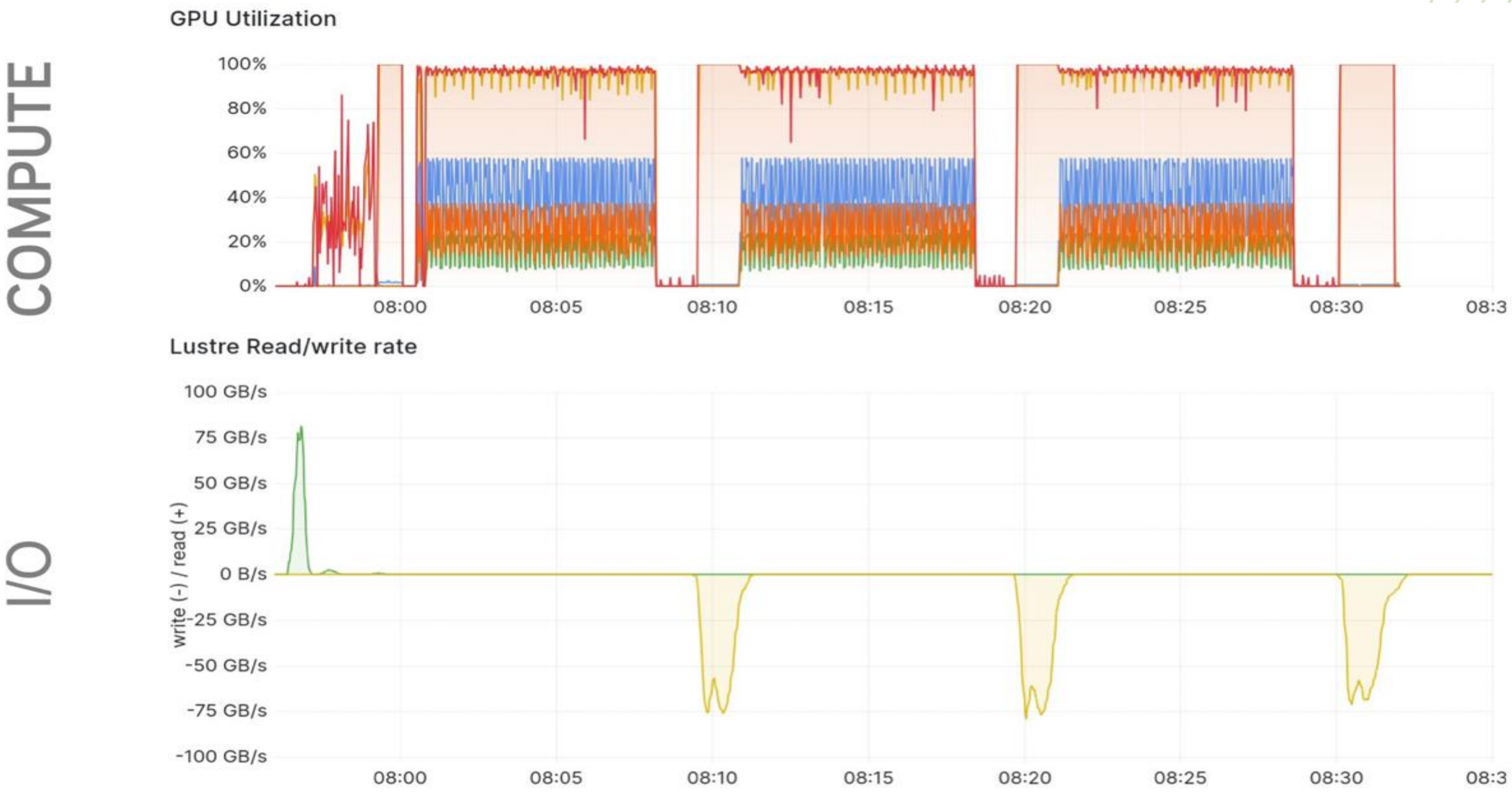
Dmitry Livshits, CEO, livshits@xinnor.io

Storage Requirements for AI workload in the Cloud

1. Virtualization
   - multiple Tenants
   - multiple file systems
2. Performance: Each tenants requires
   1. 10+ of GB/s
   2. 100K+ IOPS
3. Data resiliency
4. Low CPU consumption
5. Disaggregated composable storage



Performance requirements



More details: https://www.depts.ttu.edu/hpcc/events/LUG24/slides/Day2/LUG_2024_Talk_15-AI_Workload_Optimization_with_Lustre.pdf

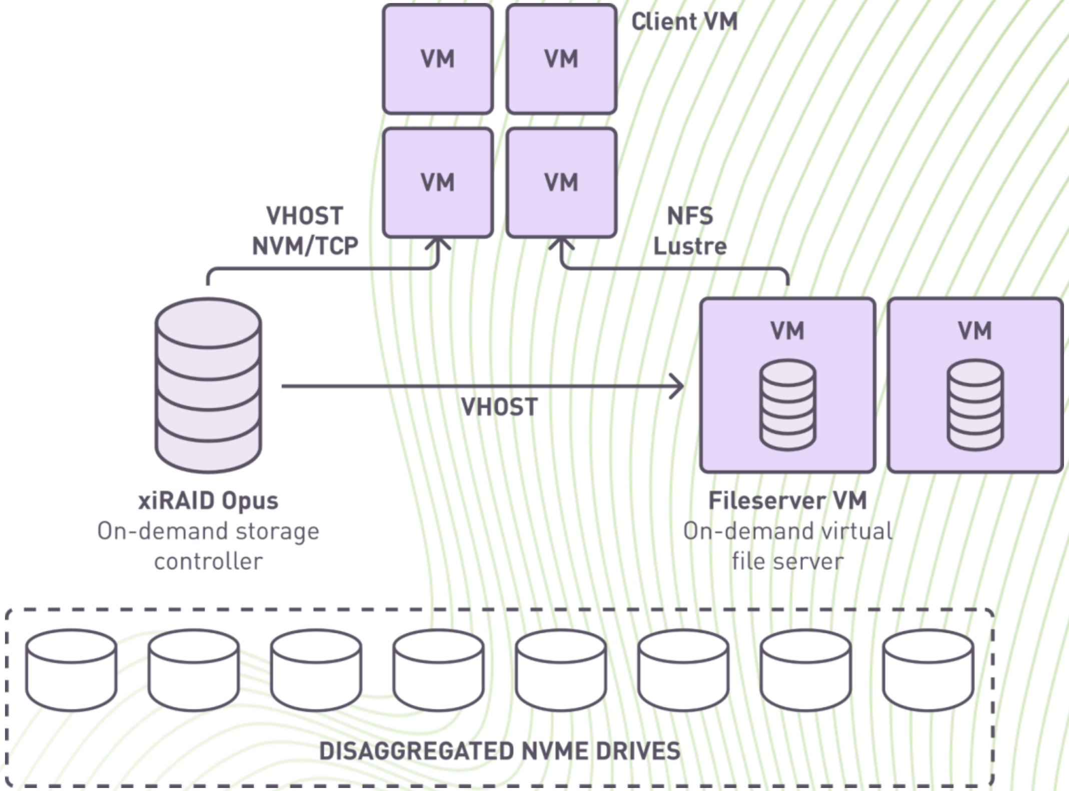
xiRAID versions

xiRAID Classic	xiRAID Opus (Optimized Performance in User Space)
operates within Linux kernel	Operates in user space, independently from the kernel
Suitable for local RAID	Suitable for network devices or virtualization
Exports a Linux block device	Can be operated via virtIO, NVMeoRDMA, NVMeoTCP
Supports all RAID levels, rebuild, and more	Additional built-in features like NVMe initiator, NVMe over TCP/RDMA, iSCSI target, and Vhost controller
managed through CLI	distributed CLI for managing multiple servers
Utilizes a small portion of all available CPU cores, distributing load evenly	Fully occupy specific dedicated CPU cores
x86 CPU only	x86 and ARM architectures (DPU)

xiSTORE Cloud = xiRAID Opus + Virtualized File Server

Advantages:

- Each tenant can deploy its own file system: PFS(Lustre/BeeGFS) or NFS
- Unbeatable performance for tenant:
  - xiRAID Opus + VHOST = 8.3M IOPs for 1 virtual volume
  - Linux kernel block device + VHOST = limited to max 250K IOPs
- Lightweight solution:
  - deployment ready just starting from 1 server node
  - + 177/30 GB/s for full stripe R/W throughput for **1 CPU core**
- DPU-ready architecture without any performance compromises

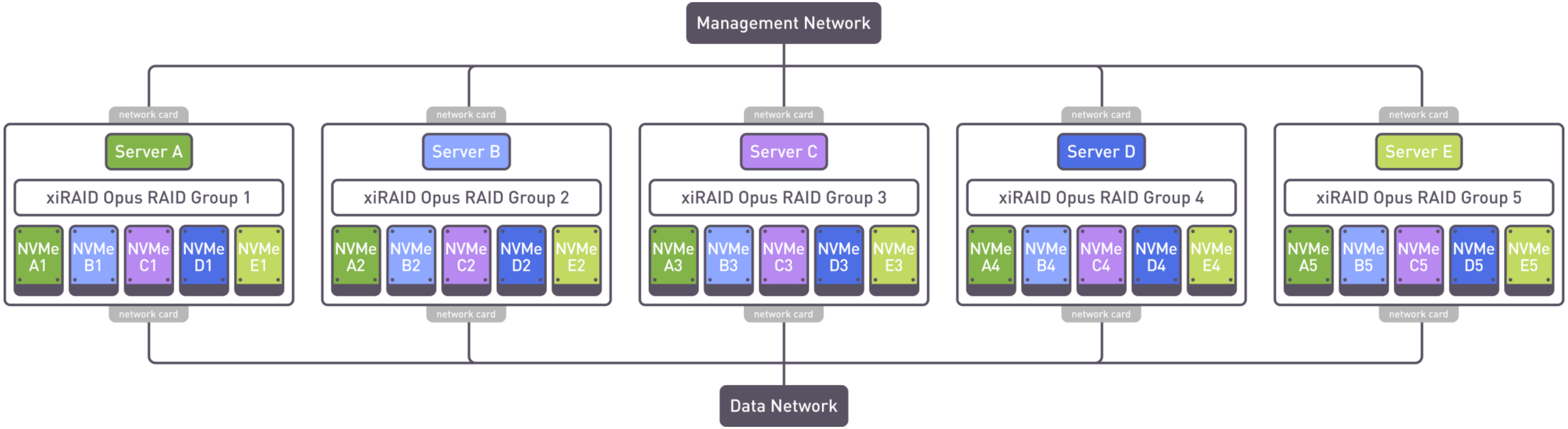

Client VM
VM
VM
VM
VM
VHOST
NVM/TCP
NFS
Lustre
VHOST
VM
VM
xiRAID Opus
On-demand storage controller
Fileserver VM
On-demand virtual file server
DISAGGREGATED NVME DRIVES

In development

xiRAID Opus in Distributed Erasure Coding for Cloud

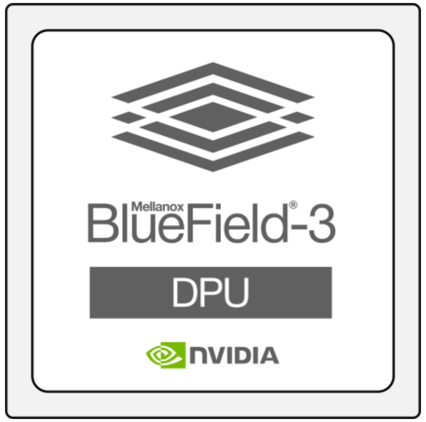
xiRAID Opus can be deployed as distributed erasure coding over multiple servers.

Each RAID group can be created by using one or more drives from each server node, to create a resilient storage solution, capable of surviving not only multiple drives failure, but also multiple node failures.



Beta release

xiRAID Opus Offloaded to BlueFiled3

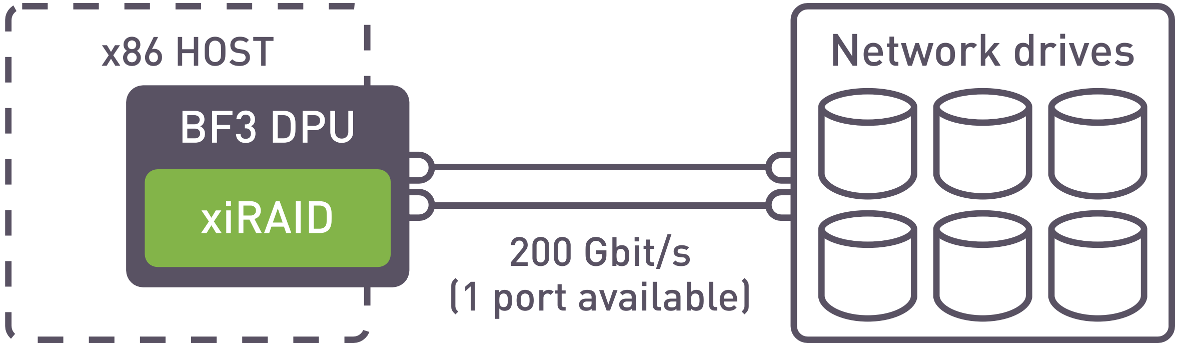


Implementation

- Network drives are visible through BF3 network 200Gbs ports
- xiRAID Opus implements the RAIDs in BF3 DPU
- The RAIDs are exposed to the host by SNAP

Advantages

- Serverless storage implementation: Zero CPU consumption
- Disaggregation: change storage capacity “on the fly” via SNAP
- Security: no need to install specialized SW or HW



xiRAID Performance on BlueField3

Tests performed using 6x Samsung PM9A3 3.84TB NVMe drives in RAID5 and 6, connected using nvme-rdma driver over IB port 200Gbit/s.

Worlkoad is running on BlueField3 (Fio plugin SPDK mode)

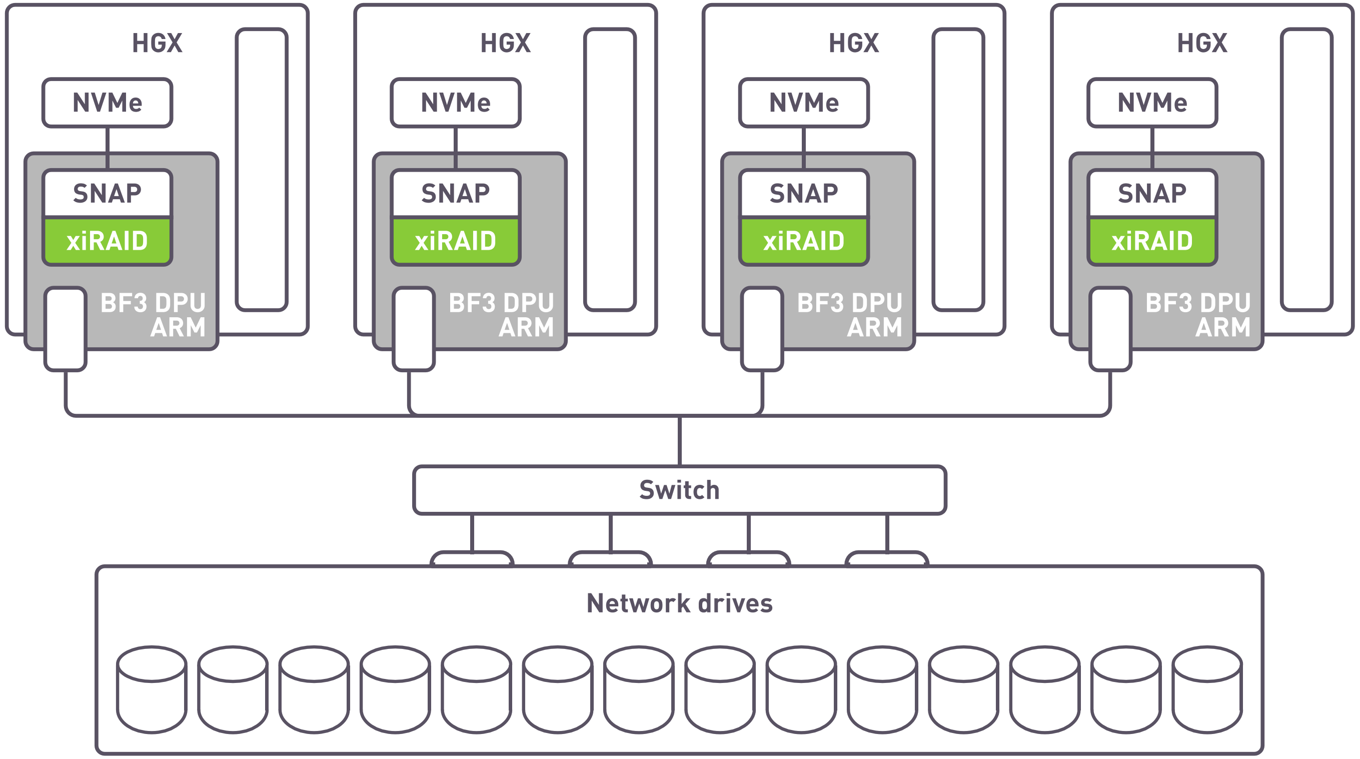
	Sequential Write (GB/s)	Sequential Read (GB/s)	Random Write (K IOPS)	Random Read (K IOPS)
Raw drives	16	24	2,064	4,080
xiRAID, RAID5	11	24	447	2,351
xiRAID, RAID6	8.2	24	328	2,352

xiRAID offloaded on BlueField3 achieved 60-100% of theoretical performance in both RAID5 and RAID6

tests run on Nvidia's lab

The use case: Server-less disaggregated storage for AI

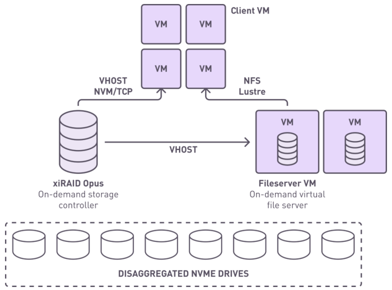
Concept



- Disaggregated storage
- Server optimization
- Network card optimization
- No Client side software
- Any host OS, any hypervisor

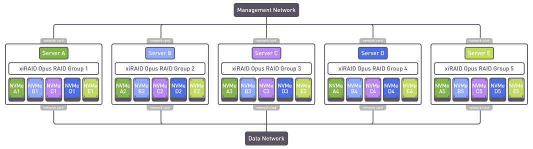
Wrap Up: cloud solution components

1 Local RAID block device + Virtualized file server



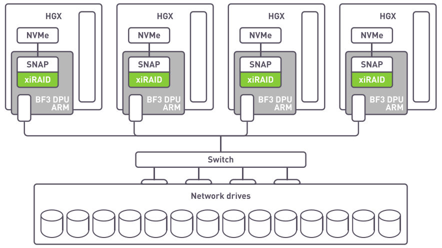
Maximize performance

2 Distributed erasure coding



Maximize resilience

3 DPU RAID Offload



Server-less implementation

Prove it yourself:
https://xinnor.io/