On Demand PFS is an appliance based preconfigured Parallel File System Storage solution built on BeeGFS, a parallel cluster file system, aimed at small to midsized HPC/AI installations.

This Document provides detailed information about components and system architecture.

On Demand PFS

White Paper

Authored by:
Rakesh Sabharwal
Founder & CEO
On Demand Systems Pte Ltd

TABLE OF CONTENTS

Executive Summary 3

Problem Statement 4

ON DEMAND PFS Solution Highlights 5

System Components 6

- HPE proliant DL Server Platform 6
- NVIDIA Mellanox HDR Switches 6
- BeeGFS Overview 7
- Xinnor xiRAID Overview 8
  - Advantages of the Xinnor Software RAID 9

ON DEMAND PFS System architecture 10

- Scalability 11
- high availability 11

Conclusion 12

REFERENCES 13

- ThinkParQ documentation 13
- xinnor documentation 13
- linbit documentation 13

On Demand systems pte ltd overview 13

- Contact Information: 14

Figure 1: xiRAID System Architecture 8

Figure 2: System Architecture (4-Nodes with IB Fabric) 10

EXECUTIVE SUMMARY

In the realm of high-performance computing (HPC) and artificial intelligence (AI), designing a storage system that strikes the right balance for efficient and optimal performance poses significant challenges. These environments, often consisting of hundreds of compute/worker nodes, handle vast amounts of data to solve intricate and complex problems. They need to support both large data sets for big data analytics and numerous small files for machine learning (ML) and AI applications without experiencing performance issues. Organizations are seeking affordable, high-performance, and highly available solutions that are also easy to manage while adapting to the growing demands of diverse I/O-intensive workloads.

The challenge is that HPC/AI storage infrastructure requirements can change quickly, requiring companies to scale up and scale out their resources. Composable Disaggregated Infrastructure (CDI) represents the modern architectural approach to data centre infrastructure, disaggregating compute, storage, and network resources into shared pools that can be composed for on-demand allocation. Composable disaggregated infrastructure (CDI) for HPC/AI storage is the key to solving this optimization problem. It enables valuable resources to be deployed through software at just the right levels and within a minimal amount of time, even inside an HPC or AI cluster.

Xinnor xiRAID is a lightweight software RAID offering which complements CDI solutions whilst performing very closely to raw device capabilities and exceeding that of traditional hardware RAID solutions.

BeeGFS is an easily deployable hardware-independent POSIX parallel file system developed with a strong focus on performance whilst designed for ease of use, simple installation, and easy management.

The purpose of this document is to showcase a combined solution of BeeGFS cluster deployment with Xinnor xiRAID volumes deployed on Commodity hardware.

PROBLEM STATEMENT

HPC environments have relied on magnetic disks as their primary storage solution for decades. Today, many applications within these environments must handle both large data blocks and numerous small files. For example, training a machine learning model may involve millions of small files, whereas big data analytics typically requires processing a single, extensive dataset. Consequently, there is a growing demand for enhanced metadata performance across various workloads. To effectively support these modern requirements, essential components of the HPC cluster—including computing, storage, and networking—must be advanced.

Many distributed parallel file systems have started to support the latest commodity compute hardware and are able to utilize high-performance NVMe SSDs effectively in HPC clusters. BeeGFS combines multiple storage servers to provide a highly scalable shared network file system using striped file contents, with which the high throughput demands of large numbers of clients can easily be satisfied.

Modern I/O-intensive workloads, such as big data, machine learning (ML), artificial intelligence (AI), and the Internet of Things (IoT), necessitate a data infrastructure that can independently scale storage and compute resources. This ensures that both components are provisioned efficiently and effectively. CDI facilitates this by organizing compute, storage, and network resources into shared pools that can be allocated on demand. As a result, compute resources can be stateless, elastic, and scalable without being tied to storage. To maintain continuous operations and services, organizations must ensure data durability and availability. Xinnor xiRAID offers a high-performance software RAID solution that supports continuous availability and fault tolerance for storage systems.

This document presents the On-Demand PFS Solution, which meets the increasing demand for high-performance parallel file systems in HPC/AI clusters. It deploys a high-performance distributed parallel file system on a commodity infrastructure. ODS's validated design for HPC/AI, featuring the BeeGFS high-capacity and high-performance storage solution with Xinnor xiRAID, provides a fully supported, user-friendly, high-throughput, scale-out parallel file system storage solution with clearly defined performance characteristics.

ON DEMAND PFS SOLUTION HIGHLIGHTS

The following sections of this paper provide an overview of On Demand PFS solution.

On Demand PFS consists of tightly coupled and pre-integrated stack using Rocky Linux, BeeGFS File System, XiRAID and Prometheus/Grafana monitoring tools. The system is built using Industry Standard Hardware, HPE Server/Storage, Nvidia Mellanox InfiniBand Switches and Adapters.

The solution combines:

- HPE's Standard Proliant Servers with low latency locally attached NVMe SSDs.
- Nvidia Mellanox IB or Ethernet network fabric.
- xiRAID is a high-performance software RAID developed specifically for NVMe storage devices to utilize up to 97% of hardware performance capabilities.
- BeeGFS is a high-performance parallel file system designed for performance-oriented environments like HPC, AI, and deep learning workloads. BeeGFS includes a distributed metadata architecture for scalability and flexibility reasons. Its most important aspect is data throughput.

By combining BeeGFS and xiRAID with HPE Standard Server and Nvidia Mellanox Fabric, organizations can benefit from the parallel file systems:

- High Performance
- High Availability
- Scalability
- Robustness
- Easy to deploy and integrate with existing infrastructure
- Easy data management
- Optimized for highly concurrent access
- BeeGFS - software with enterprise features

SYSTEM COMPONENTS

HPE PROLIANT DL SERVER PLATFORM

HPE ProLiant Gen11 servers with 4th and 5th Generation AMD EPYC™ processors drive data-intensive workloads with industry-leading performance, security, and scalability.

The solution offers a choice of three different server models based on the capacity and performance requirements.

- DL325 Gen11 – Single Socket, 1U, supports up to 10x NVMe U.3
- DL345 Gen 11 – Single Socket, 2U, supports up to 34x NVMe U.3
- DL385 Gen 11 – Dual Socket, 2U, supports up to 34x NVMe U.3

HPE and AMD have partnered to design and build some of the world’s fastest and most sustainable supercomputers. But we have also developed purpose-built, density-optimized servers that deliver unprecedented performance with sustainable cooling options.

HPE ProLiant DL345 Gen11 and HPE ProLiant DL385 Gen11 servers with AMD EPYC processors bring a new level of high-performance, scalable compute options for data management workloads. The HPE ProLiant DL345 Gen11 server offers a scalable 2U solution that delivers high-performance and flexible storage across SAS/SATA/NVMe at 1P economics.

The new HPE ProLiant DL385 Gen11 server is an accelerator-optimized 2U 2P solution that delivers exceptional compute performance, upgraded high-speed data transfer rate, and memory depth. With 4th and 5th Generation AMD EPYC processors, these HPE ProLiant Gen11 servers raise the bar for data management workload performance with next-generation 5 nm technology and support for up to 160 cores per socket (up to 320 cores for two-processor servers), translating into a more efficient database and less strain when dealing with numerous queries and multiple application requests. These new servers also offer next-level energy-efficient DDR5 memory, further boosting performance, along with the latest PCI Express 5.0 bus, doubling data transfer rates and the number of lanes, for improved bandwidth and performance. In fact, high PCIe5 lane counts deliver 4x the I/O throughput of other competitive processors.1 Moreover, extremely effective cooling solutions, lower and optimized power consumption components, and new power supply units make HPE ProLiant Gen11 servers a new low-carbon footprint standard for enterprise data centres.

NVIDIA MELLANOX HDR SWITCHES

Faster servers, high-performance storage, and increasingly complex computational applications are driving data bandwidth requirements to new heights. NVIDIA Mellanox QM8700 switches provide extremely high networking performance by delivering up to 16Tb/s of non-blocking bandwidth with extremely low latency. Static routing, adaptive routing, and advanced congestion management optimize computing efficiencies, making QM8700 ideal for top-of-rack leaf connectivity or for small to extremely large clusters.

High-performance computing (HPC) and AI environments need every last bit of the bandwidth delivered by NVIDIA Mellanox high data rate (HDR) 200Gb/s switch systems. NVIDIA QM8700 switches provide up to 40 ports of 200Gb/s full bi-directional bandwidth per port. And novel NVIDIA HDR100 technology supports up to 80 ports of 100Gb/s, enabling HDR switches to provide double-density radix for 100Gb/s data speeds, reducing the cost of network design and network topologies.

BEEGFS OVERVIEW

This storage solution is based on BeeGFS, an available source parallel file system, which offers flexibility and easy scalability. The general architecture of BeeGFS consists of four main services: management, metadata, storage, and client. The server components are implemented as user-space daemons. The client is a patchless kernel module. An additional monitoring service called Grafana is also available.

BeeGFS is a hardware-independent POSIX parallel file system (a.k.a., software-defined parallel storage) developed with a strong focus on performance and designed for ease of use, simple installation and easy management. It is designed for all performance-oriented environments, including HPC, AI, deep learning, life sciences, oil, gas, media, and entertainment.

The key elements of the BeeGFS file system are as follows:

• **MetaData Targets (MDTs)**: Stores all the metadata for the file system including filenames, permissions, time stamps, and the location of stripes of data.

• **Management Daemon (MGMTD)**: Stores management data such as configuration and all the file system components.

• **MetaData Server (MDS)**: A server that runs the metadata services.

• **Storage Targets (STs)**: Stores the data stripes of the files on a file system in the HPE Proliant DL Servers. There can be multiple storage targets in a single storage service.

• **Storage Server (SS)**: A server that runs the storage services.

• **Client Module**: The BeeGFS client kernel module is installed on the clients to allow access to data on the BeeGFS file system. To the clients, the file system appears as a single namespace that can be mounted for access.

As a parallel file system, BeeGFS stripes its files over multiple server nodes to maximize read/write performance and scalability. The server nodes work together to deliver a single file system that can be simultaneously mounted and accessed by other server nodes, commonly known as clients. These clients can see and consume the distributed file system similarly to a local file system such as NTFS, XFS, or ext4.

The four main services run on a wide range of supported Linux distributions and communicate via any TCP/IP or RDMA-capable network, including InfiniBand (IB), and RDMA over Converged Ethernet (RoCE2). The BeeGFS server services (management, storage, and metadata) are user space daemons, while the client is a native kernel module (patchless). All components can be installed or updated without rebooting, and you can run any combination of services on the same node.

XINNOR XIRAID OVERVIEW

Xinnor xiRAID ensures fast and effective access to data by allowing for the creation of high-performance RAID from NVMe and SAS/SATA SSDs. Designed for the most demanding enterprise-grade tasks, xiRAID is easy to maintain and suited for operating in large server infrastructures.

- Adjusted for the most popular Linux® distribution (Ubuntu, RHEL, Oracle® Linux, Rocky Linux, Alma Linux)
- Works with local and remote drives.
- Provides RAID as a standard Linux block device.
- POSIX API support.

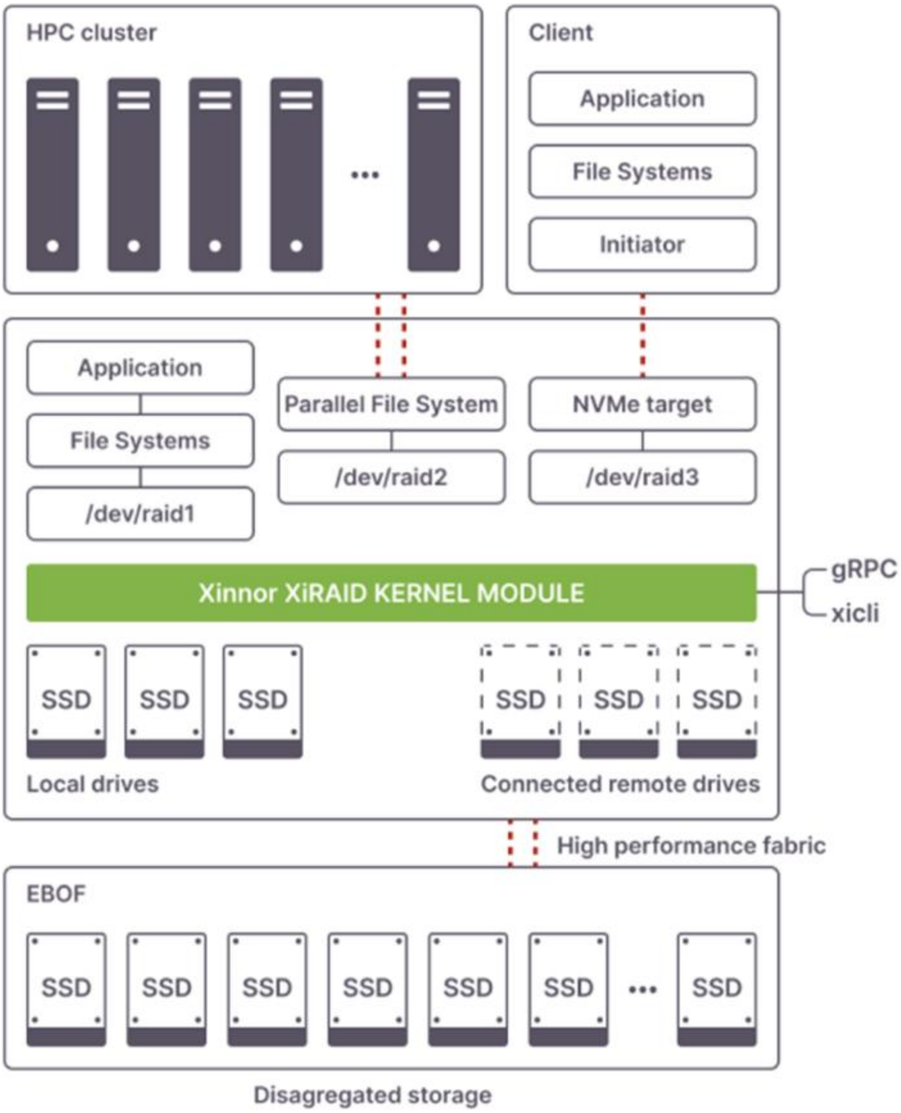

Figure 1: xiRAID System Architecture

ADVANTAGES OF THE XINNOR SOFTWARE RAID

Software RAID offers high flexibility, zero associated hardware costs and vendor-agnosticism in terms of compatibility. It’s worth noticing that software RAID is currently the only option to support the new class of NVMe-oF JBOF (EBOF) devices for disaggregated storage in the CDI world.

Due to its fast coding and decoding ability, xiRAID provides the stable performance levels needed for smooth and uninterrupted business operations. Fast RAID rebuild protects storage from extensive system downtime and mitigates the impact on workflows. This is crucial for data-intensive systems and high-density storage infrastructures where even a single drive failure can cause checksum recalculations for a vast amount of data.

1. Data Protection: xiRAID employs advanced algorithms to provide redundancy and fault tolerance, ensuring that data remains safe even in the event of multiple drive failures.

2. Performance Acceleration: In one of the deployed cluster accelerated by xiRAID was able to saturate the 2x100Gb Infiniband ports, with measured sequential read performance of 24.7GB/s. This speed allows for faster data retrieval rates, particularly beneficial for GPU-accelerated compute tasks.

3. Ease of deployment: xiRAID Classic exposes a block device within the linux kernel, for easy integration with the file system. Mounting BeeGFS over xiRAID block device is straightforward requiring minimal manual intervention, which accelerates overall deployment time.

4. Scalability: In case in the future more performance is required, it will be sufficient to add more InfiniBand cards or increase their bandwidth. Indeed, the read performance measured within the cluster exceeds 215GB/s.

4. Cost optimization: xiRAID doesn’t require any hardware. There’s no need to install a x16 PCIe card dedicated to run the RAID, so no need to waste the PCIe slot and related PCIe lanes that can be better reserved for future network expansion.

ON DEMAND PFS SYSTEM ARCHITECTURE

The unique userspace architecture concept allows users to keep the metadata access latency (e.g.,directory lookups) at a minimum and distributes the metadata across multiple servers so that each of the metadata servers stores a part of the global file system namespace.

By increasing the number of servers and disks in the system, it is possible to scale performance and capacity of the file system to the level that you need, seamlessly from small clusters up to enterprise-class systems with thousands of nodes.

Figure 1 provides a reference architecture for On Demand PFS for a 4-node cluster.

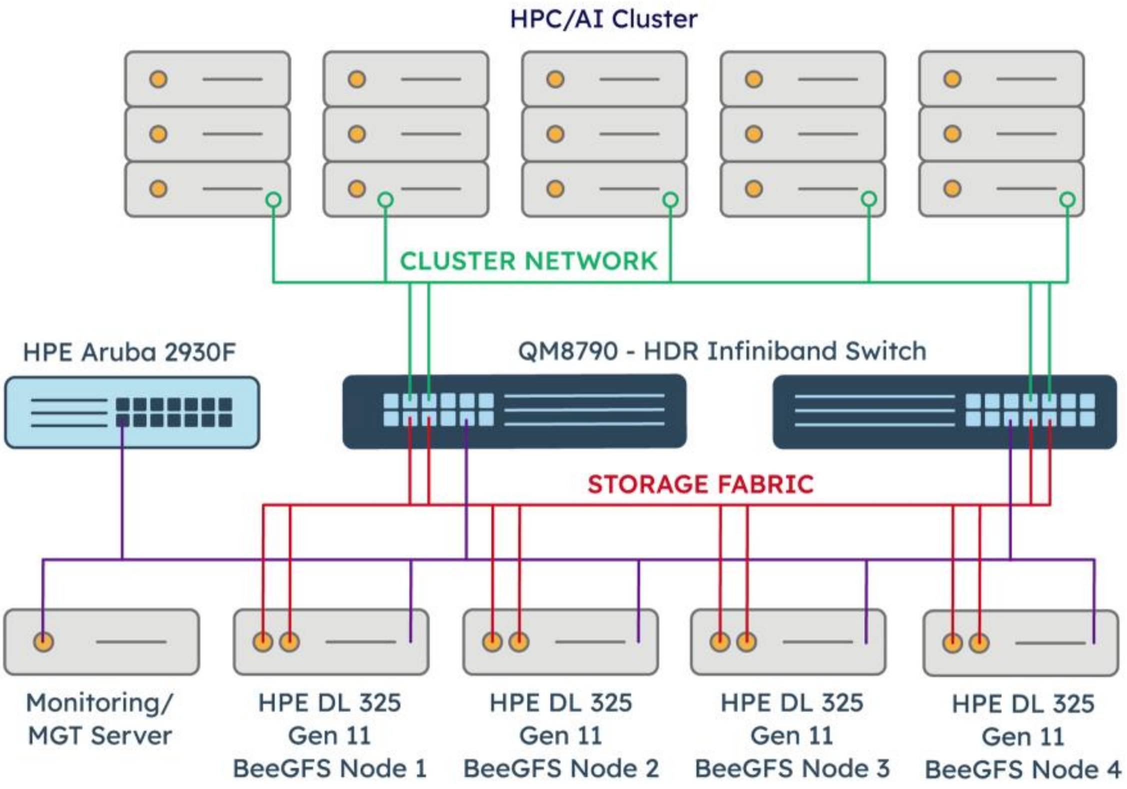
HPC/AI Cluster
CLUSTER NETWORK
HPE Aruba 2930F
QM8790 - HDR Infiniband Switch
STORAGE FABRIC
Monitoring/ MGT Server
HPE DL 325 Gen 11 BeeGFS Node 1
HPE DL 325 Gen 11 BeeGFS Node 2
HPE DL 325 Gen 11 BeeGFS Node 3
HPE DL 325 Gen 11 BeeGFS Node 4

Figure 2: System Architecture (4-Nodes with IB Fabric)

In Figure 1, the management server is an HPE DL3X5 Gen 11. The MDS function is provided by two HPE DL3X5 Gen 11 servers (node 1 and node 2). These nodes also act as Storage targets along with nodes 3 and 4.

Depending on the total usable storage requirements, we could use different HPE Server models with different sizes of NVMe disks and build the storage cluster.

SCALABILITY

The system can be horizontally scaled by adding more nodes with similar configurations. In case there is a need to add storage tiers then additional nodes with spinning disks can also be added to leverage the concept of Storage pools in BeeGFS.

Storage pools can be considered a template for the stripe pattern BeeGFS uses to assign file chunks to specific storage targets. Instead of directly assigning a stripe pattern, a user can now assign a storage pool, which is much easier and more intuitive. The storage pool assignment can be part of the directory metadata and therefore apply to all newly created files in a particular directory, or it can be part of the file metadata, applying only to that file and overriding the directory default.

HIGH AVAILABILITY

The only built-in high availability (HA) support for BeeGFS is known as Buddy Mirroring. It’s a shared-nothing architecture: Each node has access only to its own storage devices. Buddy Mirroring achieves high availability of the metadata and storage services by writing data to both a primary and secondary service and allowing reads from either. The main disadvantages of Buddy Mirroring are the requirement for double the underlying storage and a decrease in write performance.

Currently, the management service doesn’t have HA architecture. The lack of built-in high availability for the management service also creates a single point of failure that must be solved.

As part of the On Demand PFS offering, we have included HA for management service as an option. The HA is achieved by using the Distributed Replicated Storage System (DRBD) which is an open-source distributed replicated block storage software for the Linux platform and is typically used for high-performance and high availability.

ODS is also working with a few vendors dual-node and dual port NVMe technologies to further enhance the high availability options for customers.

CONCLUSION

BeeGFS on HPE Proliant Storage server with locally connected NVMe SSDs with xiRAID addresses the need of IT/HPC with a well-designed solution that is easy to manage and fully supported. The solution includes the added benefit of the open composable infrastructure environment and enables a disaggregated compute and storage platform. BeeGFS on the Storage systems shows excellent benchmark results for storage data streaming.

By leveraging the solution offered by On Demand Systems enterprises and service providers can

- Minimize the cost of the installation

- Implement a high-performance HA file system solution

- Accelerate time to market for new application services

Using this solution based on the latest components and CDI technologies, organizations can quickly deploy a proven, self-service, composable infrastructure solution, helping customers move to a more flexible, variable-cost model.

The On Demand PFS solution is also available as an OPEX model for more flexibility.

REFERENCES

The information provided in this document is based on the references below. These can be used for further advanced and in-depth reference.

THINKPARQ DOCUMENTATION

The following BeeGFS documentation from ThinkparQprovides additional and relevant information:

- BeeGFS Documentation
- General Architecture of BeeGFS File System

XINNOR DOCUMENTATION

The following xiRAID documentation provides additional and relevant information:

- xiRAID Classic Documentation

LINBIT DOCUMENTATION

The following Linbit documentation provides additional and relevant information:

- DRBD Documentation

ON DEMAND SYSTEMS PTE LTD OVERVIEW

ON DEMAND systems

ODS provides Secure High Performance Computing (HPC) and Artificial Intelligence (AI) infrastructure solutions. We provide end-to-end services from conceptualisation, design, and deployment to management. The HPC/AI solutions and services are complemented with identity and access management software and cloud-based Identity-as-a-Service that allows enterprises to securely manage identities and secure access across computer networks and cloud computing environments.

CONTACT INFORMATION:

💻 www.odspte.com

✉ info@odspte.com

☎ +65 66047271