DelftBlue Hardware

DelftBlue: The TU Delft supercomputer

DelftBlue is TU Delft's supercomputer. It is designed to meet researchers' growing need for extensive computing power to solve complex problems in physics, mechanics and dynamics. The system is characterised by flexibility, speed and user-friendliness. It has been decided to put DelftBlue into service in two phases. Phase1 went into production in June 2022 and at that time offered 11,000 CPU cores in more than 200 compute nodes.

In January 2024, with the commissioning of the phase2 hardware, DelftBlue's configuration has expanded to 17,000 CPU cores in more than 300 nodes. Also at that time, 10 new GPU nodes were added, doubling the GPU capacity to meet our users' recommendation.

Description of the DelftBlue system

The solution is built on Fujitsu hardware and makes use of: Intel processors; with a high memory throughput; InfiniBand interconnect (HDR/HDR100) technology for high throughput and low latency between all nodes; and a parallel storage subsystem, based on BeeGFS.

Compute nodes

The compute nodes are built with Intel Cascade Lake refresh processors (phase 1) and Intel Sapphire Rapids processors (phase 2) that offer high performance and energy efficiency. And deliver a combined theoretical maximum performance (Rpeak) of more than 2 PFlop/s.

The cluster consists of three different types of compute nodes:

- Standard compute nodes
- High Memory compute nodes (large memory) in 2 memory configurations
- Compute nodes with GPU's equipped with NVIDIA Tesla cards

Node types:

Node Category	Number	Cores	CPU / GPU	Memory	SSD
Compute type-a	218	48	2x Intel XEON E5-6248R 24C 3.0GHz	192 GB	480 GB
Compute type-b *	90	64	2x Intel XEON E5-6448Y 32C 2.1GHz	256 GB	480 GB
Memory type-a	6	48	2x Intel XEON E5-6248R 24C 3.0GHz	768 GB	480 GB
Memory type-b	4	48	2x Intel XEON E5-6248R 24C 3.0GHz	1,536 GB	480 GB
GPU type-a	10	48	2x AMD EPYC 7402 24C 2.80 GHz 4x NVIDIA Tesla V100S 32GB	256 GB	1.0 TB
GPU type-b *	10	64	2x Intel XEON E5-6448Y 32C 2.1GHz 4X NVIDIA Tesla A100 80GB	512 GB	1.92 TB

*added in Phase 2

Summary and performance

CPU total	Compute nodes	**338**
	CPU's	**676**
	Compute cores	**17,842**
	Rpeak (theoretical max. performance, in PFlops)	**1.45**
GPU total	GPU nodes	**20**
	GPU's	**80**
	Tensor cores	**42,880**
	CUDA cores	**481,280**
	Rpeak (theoretical max. performance, in PFlops)	**0.61**

Front-end nodes

The Cluster uses a number of front-end nodes as the entry point to the cluster for end-users and administrators.

- Login nodes
- Interactive/visualization nodes equipped with NVIDIA Quadro RTX cards

Front-end nodes:

Node Category	Number	Cores	CPU / GPU	Memory	SSD
Login	4	32	2x Intel XEON Gold E5-6226R 16C 2.9GHz	384 GB	
Interactive/visualization	2	32	2x Intel XEON Gold E5-6226R 16C 2.9GHz 1x NVIDIA Quadro RTX4000	192 GB	

Highlights / Details:

- The login nodes are meant to be the main access point for all end-users of the system. It is expected they will have a high level of competition of user sessions and therefore they have been configured with 384GB or memory. In addition, these nodes are configured as HA pairs to ensure the effects of a single node failure does not stop user access to the cluster.
- The Interactive/visualization nodes can be used for running specific interactive tasks that need to utilize a high-end graphics card for visualization or for workloads that may not be suitable for running on the cluster. Quadro RTX 4000: with 2304 CUDA cores, 288 Tensor Cores, 36 RT cores and 8 GB GDDR6 memory.
- Two File Transfer nodes are specifically included in order to provide an optimal data flow between DelftBlue and the central research storage of the TU Delft. These nodes can be used as a pre-staging or post-staging job that runs just before or after a computational job.

Interconnect

HPC applications make frequent use of communication between nodes when calculating their computational results. To maintain application efficiency even when scaling over a large number of

nodes, the interconnect must minimise overhead and enable high-speed message delivery. DelftBlue is equipped with a high performing InfiniBand network based on Mellanox InfiniBand products to build an efficient low-overhead transport fabric. This interconnect set-up not only enables efficient delivery of MPI based messages but also provides applications with high-speed access to the temporary storage area which is also accessible over the InfiniBand fabric.

Highlights/details:

- Mellanox InfiniBand HDR100/HDR interconnect configured in a Full Bisectional Bandwidth (FBB) non-blocking network fabric.
- Fat tree topology
- 100Gbps IB HDR100 HCA's per server

Storage

To enable efficient I/O throughput for computational jobs, the configuration includes a high-speed file system with 696 TB usable storage space and a throughput of at least 20 GB/s. This storage subsystem consists of:

- 6x IO servers – 2 MetaData servers and 4 Storage Servers
- 1x NetApp All flash storage/controller shelf – MetaData storage
- 4x NetApp storage/controller shelves – File data storage

Highlights/details:

- NetApp All flash subsystem for managing the Meta Data requirements
- A set of NetApp High capacity, high throughput disk subsystems providing redundancy at all levels to avoid Single Point Of Failure conditions
- High speed connectivity
- Multiple HDR100 IB links per Storage server to the IB network
- Multiple high-speed FibreChannel 32 Gbps connections between the storage servers and storage devices
- File system built using the BeeGFS parallel file system
- Dynamic Job Parallel File System "BeeOND" is a facility, which lets a user build a parallel file system from the local disks of the nodes over which their job is running. Each node of the cluster incorporates a 480GB SSD device in addition to the OS boot device. This SSD can be used to build a local parallel file system.

HPC Management solution

Bright Cluster Manger (BCM) is used for deploying and managing the cluster and HPC software environment. It includes a complete set of HPC libraries, tools, management and reporting facilities.

Operating System

Red Hat Enterprise Linux 8

Job Management

DelftBlue is equipped with Slurm job resource manager, which provides access to resources, based on the fair share principle.

End user Portal

Traditionally a HPC provides its users with SSH and FTP to access cluster resources. However, our goal is to encourage the use of the DHPC facilities by the use of a modern and easy to use portal. Therefore, DelftBlue is equipped with the Open OnDemand portal.