SVA REFERENZ-PROJEKT //
FORSCHUNGSZENTRUM JÜLICH

FORSCHUNGSZENTRUM JÜLICH OPTIMIERT HÖCHSTLEISTUNGS-RECHENZENTRUM

SVA bietet Cluster-Konzept auf Basis von SuperMicro-, NVIDIA- und Mellanox-Technologie.

HPC-CLUSTER-SYSTEM FÜR MACHINE-LEARNING- UND DEEP-LEARNING-ANWENDUNGEN

JÜLICH SUPERCOMPUTING CENTRE (JSC)

Das Forschungszentrum Jülich betreibt interdisziplinäre Spitzenforschung und entwickelt die Grundlagen für zukünftige Schlüsseltechnologien. Das Jülich Supercomputing Centre (JSC) im Forschungszentrum Jülich betreibt seit 1987 das erste deutsche Höchstleistungsrechenzentrum. Es stellt den Forschern in Deutschland und Europa über ein unabhängiges Peer-Review-Verfahren Rechenzeit der höchsten Leistungsebene zur Verfügung.

Die Forschungs- und Entwicklungsarbeiten konzentrieren sich auf mathematische Modellierung und numerische, insbesondere parallele Algorithmen für Quantenchemie, Molekulardynamik und Monte-Carlo-Simulationen. In der Informatik liegt der Schwerpunkt auf den Gebieten Cluster-Computing, Leistungsanalyse paralleler Programme, Visualisierung, Quanten-Computing und föderierte Daten-Services.

HERAUSFORDERUNG

Das Jülich Supercomputing Centre (JSC) plante die Beschaffung eines neuen Cluster-Systems, sowohl als Test- und Development-System für Deep-Learning-Algorithmen und entsprechenden Tools als auch als kleines Produktionssystem für Machine-Learning, Deep Learning, Data Analytics und HPC. Zu diesem Zweck sollte eine, dem vorhandenen Budget angepasste, performante, gut ausbalancierte sowie auf dem neuesten Stand der Technik basierende Lösung erarbeitet und angeboten werden.

Dies galt sowohl für die Anbindung des schon vorhandenen Storage als auch für den High Speed Interconnect zwischen den Knoten. Der High Speed Interconnect sollte dabei hoch parallelisiertes Arbeiten mit verteilten modernen GPUs möglichst effizient unterstützen, um eine maximale Performance zu erreichen. Die Storage-Anbindung musste sich problem-

TECHNIK- UND SERVICE-ANFORDERUNGEN

los in die vorhandene Infrastruktur des JSC einfügen lassen. Des Weiteren waren Anforderungen an den Platzbedarf für die Infrastruktur und eine einfache Portierung der Workloads von und zu den existierenden Supercomputing Cluster zu berücksichtigen, die keine eher exotischen ML/AI Accelerators (GPUs) zuließen.

Neben den technischen Anforderungen galt es auch, verschiedene Service-Anforderungen des JSC zu erfüllen. Dazu gehörte ein Service-Konzept für die Komponenten der unterschiedlichen Partner, der betriebsbereite Aufbau des Clusters sowie Performance-Messungen. Darüber hinaus sollte ein Konzept für die sichere und garantierte Vernichtung von Daten auf ausgetauschten defekten Speichermedien (SSDs) erstellt werden.

SVA-LÖSUNG: CLUSTER-KONZEPT MIT SUPERMICRO, NVIDIA UND MELLANOX

SVA schlug ein Cluster-Konzept mit zukunftsweisender Technologie der Partner SuperMicro, NVIDIA und Mellanox mit einer hoch-performanten Architektur vor. Die Dual Socket HPC Server von SuperMicro verfügen über die bestmögliche Ausstattung für die Anforderungen – vier NVIDIA Volta100 GPU auf SXM2 Sockets, die über PCIe-Switches an die CPUs angeschlossen sind, und vier zusätzlichen PCIe x 16 Slots für Infiniband und Ethernet Cards.

Mit den zwei Single Port Infiniband Cards in den Servern kann das Feature „GPUDIRECT“ von NVIDIA/Mellanox optimal betrieben werden. Eine zusätzliche 100 Gbit Ethernet Card ermöglicht den Compute Nodes einen hochperformanten Anschluss an das Storage-Netzwerk. Separate interne NVMe SSDs für das Operating System und Scratch Storage runden die optimale Architektur ab. Der Einsatz der NVIDIA v100 GPU mit jeweils 32 GB HB2 Memory und die Verwendung des gegenüber PCIe schnelleren NVLINKs von NVIDIA ermöglicht eine maximale Performance für alle Berechnungen.

MAXIMALE PERFORMANCE MIT NVIDIA

Der einzelne Compute Node verfügt insgesamt über 4 x PCIe 3.0 x 16 Slots und 2 x PCIe 3.0 x 16 Switches und 2 x PCIe 3.0 x 16 Bridges zu den SXM2 Sockets. Dem Anwender steht nach der Bestückung mit Controllern, NVMe Devices und den GPUs noch ein PCIe 3.0 x 16 Slot zur Verfügung. Ersichtlich ist die Optimierung des GPUDIRECT über die beiden Infiniband HCA.

Um Interferenzen in der Performance durch die verschiedenen IO-Profile des Storage Traffics und des MPI/GPU Traffics zu verhindern, sollten trotz höherer Kosten Interconnect-Netzwerk und Storage-Netzwerk getrennt werden. Dadurch wurde dem JSC ein Interconnect-Netzwerk auf Basis von Infiniband EDR (100 Gbit/s) und ein separates Storage-Netzwerk auf Basis von 100 Gbit/s Ethernet angeboten, was optimale Performance und Latency für beide Anwendungen ermöglicht. Die Bandbreite des Storage-Netzwerkes war z. B. insgesamt mit 1700 GB/s bei geforderten minimalen 500 GB/s überragend.

JÜLICH Forschungszentrum

SUPERMICRO

NVIDIA

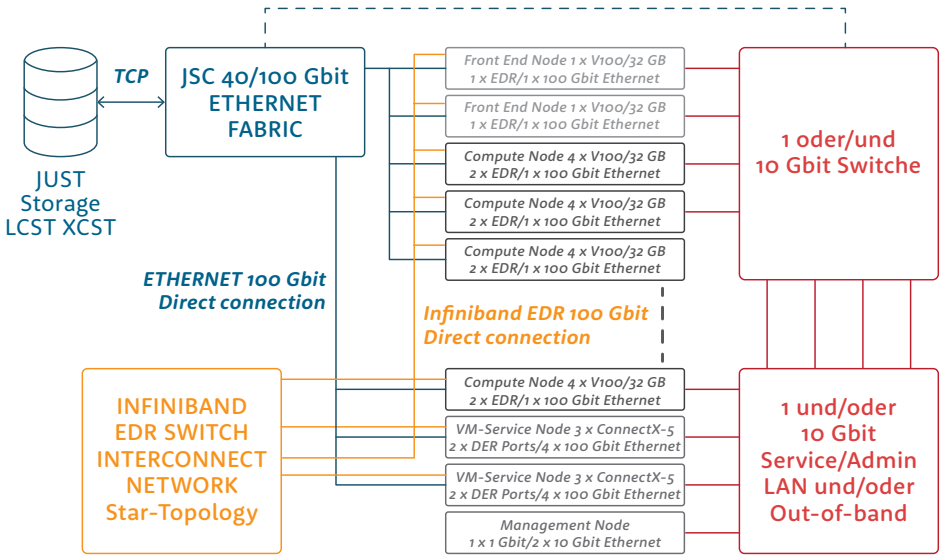
SERVICE-KONZEPT

Das Service-Konzept wurde gemäß der JSC-Vorgabe von „Hardware-Austausch und -Diagnose am nächsten Tag“ (Next Business Day) über einen Zeitraum von fünf Jahren plus optional einem Jahr zusätzlich erstellt. Diagnose und Hardware-Austausch werden durch den Hersteller SuperMicro durchgeführt, wobei die SVA-Experten immer informiert und im Problemfall selbst aktiv werden. Hinzu kommt Digital Media Retention (DMR) für Non-Volatile Devices nach BSI, durchgeführt von einem externen, spezialisierten und von SVA beauftragten Unternehmen. Somit wird eine möglichst geringe Involvierung der JSC-Mitarbeiter bei einem Service-Fall ermöglicht.

SAUBERES PROJEKTMANAGEMENT

FAZIT

Das Cluster-Konzept mit aktuellster Hardware konnte überzeugen: Die Performance wurde im Abnahmetest durch High Performance Linpack (HPL) und IOR Benchmarks bestätigt. Schlüssel für die erfolgreiche Umsetzung des Projekts waren nicht nur ein sauberes Projektmanagement, sondern auch eine offene und direkte Kommunikation zwischen JSC und SVA, die bei auftretenden Problemen schnell und konsequent eine Lösung ermöglicht hat.



AUF EINEN BLICK

AUFGABE

Konzeption eines HPC-Cluster-Systems primär für Machine-Learning- und Deep-Learning-Anwendungen

SYSTEME UND SOFTWARE

15 x Compute Nodes:
SuperMicro SuperServer SYS-1029GQ-TVRT
> 2 x Intel® Xeon® Gold 6126 Processor, 12 Cores, 2,6 – 3,7 GHz, Cache 19,25 MB, 3 x UPI
> Memory 192 GB DDR4 Dimm
> 4 x NVIDIA Tesla V100 SXM2, 32 GB HBM2, NVLINK
> NVMe device M.2 und 2 x 4 TB Intel NVMe SSD
> 3 x Mellanox ConnectX-5 VPI, EDR IB und 100 Gbit Ethernet single port HBA

2 x Front End Nodes (Login Server):
SuperMicro SuperServer SYS-1029GP-TR
> 2 x Intel® Xeon® Gold 6126 Processor, 12 Cores, 2,6 – 3,7 GHz, Cache 19,25 MB, 3 x UPI
> Memory 192 GB DDR4 Dimm
> 1 x NVIDIA Tesla PClex16, 32 GB HBM2, NVLINK
> 2 x 960 GB Intel Sata Disks
> 2 x Mellanox ConnectX-5 VPI, EDR IB und 100 Gbit Ethernet

3 x Service Nodes (VMware Systeme):
SuperMicro SuperServer SYS-6029P-TRT
> 2 x Intel® Xeon® Gold, 6130, 16 Cores, 2,1 GHz, Cache 22 MB, 3 x UPI
> Memory 192 GB DDR4 Dimm
> 2 x 6 TB Intel Sata Disks
> 3 x Mellanox ConnectX-5 VPI, EDR IB und 100 Gbit Ethernet dual port HBA

1 x Management Node:
SuperMicro SYS-5019P-WT
> 1 x Intel® Xeon® Gold, 6138, 20 Cores, 2 GHz, Cache 27,5 MB, 2 x UPI
> Memory 192 GB DDR4 Dimm
> 2 x 6 TB Intel Sata Disks
> 3 x Mellanox ConnectX-5 VPI, EDR IB und 100 Gbit Ethernet dual port HBA

1 x Mellanox 78xx Infiniband EDR Switch

VORTEILE

> Optimale, State-of-the-Art Architektur für GPU's und PCIe Connections
> Energieeffiziente und platzsparende Systeme
> High Speed 100 Gbit Infiniband und Ethernet-Netzwerk RDMA Connections

KONTAKT
SVA System Vertrieb
Alexander GmbH
Borsigstraße 14
65205 Wiesbaden
Tel. +49 6122 536-0
Fax +49 6122 536-399
mail@sva.de
www.sva.de