AI Converged Infrastructures

NetApp Solutions

NetApp
October 28, 2022

This PDF was generated from https://docs.netapp.com/us-en/netapp-solutions/ai/nva-1151-design-link.html on October 28, 2022. Always check docs.netapp.com for the latest.

Table of Contents

AI Converged Infrastructures . . . 1
NetApp ONTAP AI with NVIDIA . . . 1
NetApp EF-Series AI with NVIDIA . . . 2
TR-4859: Deploying IBM Spectrum Scale with NetApp E-Series Storage - Installation and validation . . . 3
TR-4810: NetApp ONTAP and Lenovo ThinkSystem SR670 for AI and ML Model Training Workloads . . . 3
TR-4815: NetApp AFF A800 and Fujitsu Server PRIMERGY GX2570 M5 for AI and ML Model Training Workloads . . . 3

AI Converged Infrastructures

NetApp ONTAP AI with NVIDIA

Overview of ONTAP AI converged infrastructure solutions from NetApp and NVIDIA.

NetApp ONTAP AI with NVIDIA DGX A100 Systems

- Design Guide
- Deployment Guide

NetApp ONTAP AI with NVIDIA DGX A100 Systems and Mellanox Spectrum Ethernet Switches

- Design Guide
- Deployment Guide

NVA-1151-DESIGN: NetApp ONTAP AI with NVIDIA DGX A100 Systems Design Guide

David Arnette and Sung-Han Lin, NetApp

NVA-1151-DESIGN describes a NetApp Verified Architecture for machine learning and artificial intelligence workloads using NetApp AFF A800 storage systems, NVIDIA DGX A100 systems, and NVIDIA Mellanox network switches. It also includes benchmark test results for the architecture as implemented.

https://www.netapp.com/pdf.html?item=/media/19432-nva-1151-design.pdf

NVA-1151-DEPLOY: NetApp ONTAP AI with NVIDIA DGX A100 Systems

David Arnette, NetApp

NVA-1151-DEPLOY includes storage system deployment instructions for a NetApp Verified Architecture (NVA) for machine learning (ML) and artificial intelligence (AI) workloads using NetApp AFF A800 storage systems, NVIDIA DGX A100 systems, and NVIDIA Mellanox network switches. It also includes instructions for running validation benchmark tests after deployment is complete.

https://www.netapp.com/pdf.html?item=/media/20708-nva-1151-deploy.pdf

NVA-1153-DESIGN: NetApp ONTAP AI with NVIDIA DGX A100 Systems and Mellanox Spectrum Ethernet Switches

David Arnette and Sung-Han Lin, NetApp

NVA-1153-DESIGN describes a NetApp Verified Architecture for machine learning (ML) and artificial intelligence (AI) workloads using NetApp AFF A800 storage systems, NVIDIA DGX A100 systems, and NVIDIA Mellanox Spectrum SN3700V 200Gb Ethernet switches. This design features RDMA over Converged Ethernet (RoCE) for the compute cluster interconnect fabric to provide customers with a completely ethernet-based architecture for high-performance workloads. This document also includes benchmark test results for the architecture as implemented.

https://www.netapp.com/pdf.html?item=/media/21793-nva-1153-design.pdf

NVA-1153-DEPLOY: NetApp ONTAP AI with NVIDIA DGX A100 ystems and Mellanox Spectrum Ethernet Switches

David Arnette, NetApp

NVA-1153-DEPLOY includes storage-system deployment instructions for a NetApp Verified Architecture for machine learning (ML) and artificial intelligence (AI) workloads using NetApp AFF A800 storage systems, NVIDIA DGX A100 systems, and NVIDIA Mellanox Spectrum SN3700V 200Gb Ethernet switches. It also includes instructions for executing validation benchmark tests after deployment is complete.

https://www.netapp.com/pdf.html?item=/media/21789-nva-1153-deploy.pdf

NetApp EF-Series AI with NVIDIA

Overview of EF-Series AI converged infrastructure solutions from NetApp and NVIDIA.

EF-Series AI with NVIDIA DGX A100 Systems and BeeGFS

- Design Guide
- Deployment Guide
- BeeGFS Deployment Guide

NVA-1156-DESIGN: NetApp EF-Series AI with NVIDIA DGX A100 Systems and BeeGFS

Abdel Sadek, Tim Chau, Joe McCormick and David Arnette, NetApp

NVA-1156-DESIGN describes a NetApp Verified Architecture for machine learning (ML) and artificial intelligence (AI) workloads using NetApp EF600 NVMe storage systems, the BeeGFS parallel file system, NVIDIA DGX A100 systems, and NVIDIA Mellanox Quantum QM8700 200Gbps IB switches. This design features 200Gbps InfiniBand (IB) for the storage and compute cluster interconnect fabric to provide customers with a completely IB-based architecture for high-performance workloads. This document also includes benchmark test results for the architecture as implemented.

https://www.netapp.com/pdf.html?item=/media/25445-nva-1156-design.pdf

NVA-1156-DEPLOY: NetApp EF-Series AI with NVIDIA DGX A100 Systems and BeeGFS

Abdel Sadek, Tim Chau, Joe McCormick, and David Arnette, NetApp

This document describes a NetApp Verified Architecture for machine learning (ML) and artificial intelligence (AI) workloads using NetApp EF600 NVMe storage systems, the ThinkParQ BeeGFS parallel file system, NVIDIA DGX A100 systems, and NVIDIA Mellanox Quantum QM8700 200Gbps InfiniBand (IB) switches. This document also includes instructions for executing validation benchmark tests after the deployment is complete.

https://www.netapp.com/pdf.html?item=/media/25574-nva-1156-deploy.pdf

TR-4859: Deploying IBM Spectrum Scale with NetApp E-Series Storage - Installation and validation

Chris Seirer, NetApp

TR-4859 describes the process of deploying a full parallel file system solution based on IBM's Spectrum Scale software stack. TR-4859 is designed to provide details on how to install Spectrum Scale, validate the infrastructure, and manage the configuration.

https://www.netapp.com/pdf.html?item=/media/22029-tr-4859.pdf

TR-4810: NetApp ONTAP and Lenovo ThinkSystem SR670 for AI and ML Model Training Workloads

Karthikeyan Nagalingam, NetApp
Miroslav Hodak, Lenovo

TR-4810 describes a cost-effective, entry-level compute and storage architecture to deploy GPU-based artificial intelligence (AI) training on NetApp storage controllers and Lenovo ThinkSystem servers. The setup is designed as a shared resource for small to medium-sized teams running multiple training jobs in parallel.

TR-4810 provides performance data for the industry-standard MLPerf benchmark evaluating image classification training with TensorFlow on V100 GPUs. To measure performance, we used ResNet50 with the ImageNet dataset, a batch size of 512, half precision, CUDA, and cuDNN. We performed this analysis using four-GPU SR670 servers and an entry-level NetApp storage system. The results show highly efficient performance across the multiple use cases tested here—shared, multiuser, multijob cases, with individual jobs scaling up to four servers. Large scale-out jobs were less efficient but still feasible

https://www.netapp.com/media/17115-tr-4810.pdf

TR-4815: NetApp AFF A800 and Fujitsu Server PRIMERGY GX2570 M5 for AI and ML Model Training Workloads

David Arnette, NetApp
Takashi Oishi, Fujitsu

This solution focuses on a scale-out architecture to deploy artificial intelligence systems with NetApp storage systems and Fujitsu servers. The solution was validated with MLperf v0.6 model-training benchmarks using Fujitsu GX2570 servers and a NetApp AFF A800 storage system.

https://www.netapp.com/pdf.html?item=/media/17215-tr4815.pdf

Copyright information

Copyright © 2022 NetApp, Inc. All Rights Reserved. Printed in the U.S. No part of this document covered by copyright may be reproduced in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or storage in an electronic retrieval system—without prior written permission of the copyright owner.

Software derived from copyrighted NetApp material is subject to the following license and disclaimer:

THIS SOFTWARE IS PROVIDED BY NETAPP “AS IS” AND WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, WHICH ARE HEREBY DISCLAIMED. IN NO EVENT SHALL NETAPP BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

NetApp reserves the right to change any products described herein at any time, and without notice. NetApp assumes no responsibility or liability arising from the use of products described herein, except as expressly agreed to in writing by NetApp. The use or purchase of this product does not convey a license under any patent rights, trademark rights, or any other intellectual property rights of NetApp.

The product described in this manual may be protected by one or more U.S. patents, foreign patents, or pending applications.

LIMITED RIGHTS LEGEND: Use, duplication, or disclosure by the government is subject to restrictions as set forth in subparagraph (b)(3) of the Rights in Technical Data -Noncommercial Items at DFARS 252.227-7013 (FEB 2014) and FAR 52.227-19 (DEC 2007).

Data contained herein pertains to a commercial product and/or commercial service (as defined in FAR 2.101) and is proprietary to NetApp, Inc. All NetApp technical data and computer software provided under this Agreement is commercial in nature and developed solely at private expense. The U.S. Government has a non-exclusive, non-transferrable, nonsublicensable, worldwide, limited irrevocable license to use the Data only in connection with and in support of the U.S. Government contract under which the Data was delivered. Except as provided herein, the Data may not be used, disclosed, reproduced, modified, performed, or displayed without the prior written approval of NetApp, Inc. United States Government license rights for the Department of Defense are limited to those rights identified in DFARS clause 252.227-7015(b) (FEB 2014).

Trademark information

NETAPP, the NETAPP logo, and the marks listed at http://www.netapp.com/TM are trademarks of NetApp, Inc. Other company and product names may be trademarks of their respective owners.