IEEE SA
STANDARDS ASSOCIATION

INDUSTRY CONNECTIONS REPORT

IEEE 802 Nendica Report:

Intelligent Lossless Data Center Networks

IEEE

3 Park Avenue | New York, NY 10016-5997 | USA

Documents
Document Sharing Hub

Full document details: IEEE 802 Nendica Report: Intelligent Lossless Data Center Networks
Limited to: Shanghai Jiaotong University. Downloaded on June 18,2021 at 02:08:55 UTC from IEEE Xplore. Restrictions apply.

IEEE 802 Nendica Report: Intelligent Lossless Data Center Networks

Editor

Name	Affiliation
Guo, Liang	CAICT/ODCC
Congdon, Paul	Huawei

Nendica Chair

Name	Affiliation
Marks, Roger	Huawei

Contributors/Supporters

Name	Affiliation
Li, Jie	CAICT/ODCC
Gao, Feng	Baidu
Gu, Rong	China Mobile
Zhao, Jizhuang	China Telecom
Chen, Chuansheng	Tencent
Yin, Yue	Huawei
Song, Qingchun	Nvidia
Liu, Jun	Cisco
He, Zongying	Broadcom
Sun, Liyang	Huawei
Tang, Guangming	Meituan
Quan, Hao	Meituan
Tao, Chunlei	JD
Wang, Shaopeng	CAICT/ODCC

Copyright © 2021 IEEE. All rights reserved.
Documents Document Sharing Hub
Full document Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 18,2021 at 02:08:55 UTC from IEEE Xplore. Restrictions apply.

Trademarks and Disclaimers

IEEE believes the information in this publication is accurate as of its publication date; such information is subject to change without notice. IEEE is not responsible for any inadvertent errors.

Copyright © 2021 IEEE. All rights reserved.

IEEE owns the copyright to this Work in all forms of media. Copyright in the content retrieved, displayed, or output from this Work is owned by IEEE and is protected by the copyright laws of the United States and by international treaties. IEEE reserves all rights not expressly granted.

IEEE is providing the Work to you at no charge. However, the Work is not to be considered within the "Public Domain," as IEEE is, and at all times shall remain the sole copyright holder in the Work.

Except as allowed by the copyright laws of the United States of America or applicable international treaties, you may not further copy, prepare, and/or distribute copies of the Work, nor significant portions of the Work, in any form, without prior written permission from IEEE.

Requests for permission to reprint the Work, in whole or in part, or requests for a license to reproduce and/or distribute the Work, in any form, must be submitted via email to stds-ipr@ieee.org, or in writing to:

IEEE SA Licensing and Contracts
445 Hoes Lane
Piscataway, NJ 08854

Comments on this report are welcomed by Nendica: the IEEE 802 "Network Enhancements for the Next Decade" Industry Connections Activity: <https://1.ieee802.org/802-nendica>

Comment submission instructions are available at: <https://1.ieee802.org/802-nendica/nendica-dcn>

The Institute of Electrical and Electronics Engineers, Inc.
3 Park Avenue, New York, NY 10016-5997, USA

Copyright © 2021 by The Institute of Electrical and Electronics Engineers, Inc.
All rights reserved. Published June 2021. Printed in the United States of America.

IEEE and 802 are registered trademarks in the U.S. Patent & Trademark Office, owned by The Institute of Electrical and Electronics Engineers, Incorporated.

PDF: ISBN 978-1-5044-7741-3 STDVA24816

IEEE prohibits discrimination, harassment, and bullying. For more information, visit http://www.ieee.org/web/aboutus/whatis/policies/p9-26.html.

No part of this publication may be reproduced in any form, in an electronic retrieval system, or otherwise, without the prior written permission of the publisher.

To order IEEE Press Publications, call 1-800-678-IEEE.
Find IEEE standards and standards-related product listings at: http://standards.ieee.org

Copyright © 2021 IEEE. All rights reserved.

Documents
Document Sharing Hub

Full document Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 18,2021 at 02:08:55 UTC from IEEE Xplore. Restrictions apply.

NOTICE AND DISCLAIMER OF LIABILITY CONCERNING THE USE OF IEEE SA INDUSTRY CONNECTIONS DOCUMENTS

This IEEE Standards Association ("IEEE SA") Industry Connections publication ("Work") is not a consensus standard document. Specifically, this document is NOT AN IEEE STANDARD. Information contained in this Work has been created by, or obtained from, sources believed to be reliable, and reviewed by members of the IEEE SA Industry Connections activity that produced this Work. IEEE and the IEEE SA Industry Connections activity members expressly disclaim all warranties (express, implied, and statutory) related to this Work, including, but not limited to, the warranties of: merchantability; fitness for a particular purpose; non-infringement; quality, accuracy, effectiveness, currency, or completeness of the Work or content within the Work. In addition, IEEE and the IEEE SA Industry Connections activity members disclaim any and all conditions relating to: results; and workmanlike effort. This IEEE SA Industry Connections document is supplied "AS IS" and "WITH ALL FAULTS."

Although the IEEE SA Industry Connections activity members who have created this Work believe that the information and guidance given in this Work serve as an enhancement to users, all persons must rely upon their own skill and judgment when making use of it. IN NO EVENT SHALL IEEE OR IEEE SA INDUSTRY CONNECTIONS ACTIVITY MEMBERS BE LIABLE FOR ANY ERRORS OR OMISSIONS OR 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 WORK, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE AND REGARDLESS OF WHETHER SUCH DAMAGE WAS FORESEEABLE.

Further, information contained in this Work may be protected by intellectual property rights held by third parties or organizations, and the use of this information may require the user to negotiate with any such rights holders in order to legally acquire the rights to do so, and such rights holders may refuse to grant such rights. Attention is also called to the possibility that implementation of any or all of this Work may require use of subject matter covered by patent rights. By publication of this Work, no position is taken by the IEEE with respect to the existence or validity of any patent rights in connection therewith. The IEEE is not responsible for identifying patent rights for which a license may be required, or for conducting inquiries into the legal validity or scope of patents claims. Users are expressly advised that determination of the validity of any patent rights, and the risk of infringement of such rights, is entirely their own responsibility. No commitment to grant licenses under patent rights on a reasonable or non-discriminatory basis has been sought or received from any rights holder. The policies and procedures under which this document was created can be viewed at http://standards.ieee.org/about/sasb/iccom/.

This Work is published with the understanding that IEEE and the IEEE SA Industry Connections activity members are supplying information through this Work, not attempting to render engineering or other professional services. If such services are required, the assistance of an appropriate professional should be sought. IEEE is not responsible for the statements and opinions advanced in this Work.

Copyright © 2021 IEEE. All rights reserved.
Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 18,2021 at 02:08:55 UTC from IEEE Xplore. Restrictions apply.
Full document Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 18,2021

TABLE OF CONTENTS

Editor 1

Nendica Chair 1

Contributors/Supporters 1

1 INTRODUCTION 5

Scope 5

Purpose 5

2 BRINGING THE DATA CENTER TO LIFE 5

A new world with data everywhere 5

3 EVOLVING DATA CENTER REQUIREMENTS AND TECHNOLOGY 7

Previous Data Center Bridging Standards 7

Requirements evolution 8

Characteristics of AI computing 9

Evolving technologies 11

4 CHALLENGES WITH TODAY'S DATA CENTER NETWORK 21

High throughput and low latency tradeoff 21

Deadlock free lossless network 23

Congestion control issues in large-scale data center networks 24

Configuration complexity of congestion control algorithms 26

5 NEW TECHNOLOGIES TO ADDRESS NEW DATA CENTER PROBLEMS 28

Hybrid transports for low latency and high throughput 28

PFC deadlock prevention using topology recognition 29

Improving Congestion Notification 32

Addressing configuration complexity 33

6 STANDARDIZATION CONSIDERATIONS 36

7 CONCLUSION 38

8 CITATIONS 38

Documents
Document Sharing Hub

Full document details: IEEE 802 Nendica Report: Intelligent Lossless Data Center Networks 2021

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 18,2021 at 02:08:55 UTC from IEEE Xplore. Restrictions apply.

1 Introduction

This paper is the result of a work item [19] within the IEEE 802 "Network Enhancements for the Next Decade" Industry Connections Activity known as Nendica.[1] The paper extends a previous report, "IEEE 802 Nendica Report: The Lossless Network for Data Centers" [20] published on August 17, 2018. This update provides additional background on evolving use cases in modern data centers and proposes solutions to additional problems identified by this report.

Scope

The scope of this report is the exploration of networking technologies to support the requirements of modern data center networks that include support for high performance computing and artificial intelligence applications. Solutions to address challenges created by evolving requirements and new age technologies are proposed. Standardization considerations are identified.

Purpose

The purpose of this report is to frame high-level solutions to issues and challenges with modern data center Networks. The report includes background and technical analyses of current data center environments as they are applied to the evolving needs of target applications. The report highlights new technologies that are changing the dynamics and operation of the data center Network. The results of the analysis lead to identification and recommendation of future standardization activities.

2 Bringing the data center to life

A new world with data everywhere

Digital transformation is driving change in both our personal and professional lives. Workflows and personal interactions are turning to digital processes and automated tools that are enabled by the Cloud, Mobility, and the Internet of Things. The intelligence behind the digital transformation is Artificial Intelligence (AI). Data centers running AI applications with massive amounts of data are recasting that data into pertinent information, automated human interactions, and refined decision making (see Figure 1). The need to interact with the data center in real-time is more important than ever in today's world where augmented reality, voice recognition, and contextual searching demand

[1]Numbers within brackets refer to citation references found in Section 8.

Copyright © 2021 IEEE. All rights reserved.

Documents Document Sharing Hub

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 18,2021 at 02:08:55 UTC from IEEE Xplore. Restrictions apply.

Full document Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 18,2021

immediate results. Data center networks must deliver unprecedented levels of performance, scale, and reliability to meet these real-time demands.

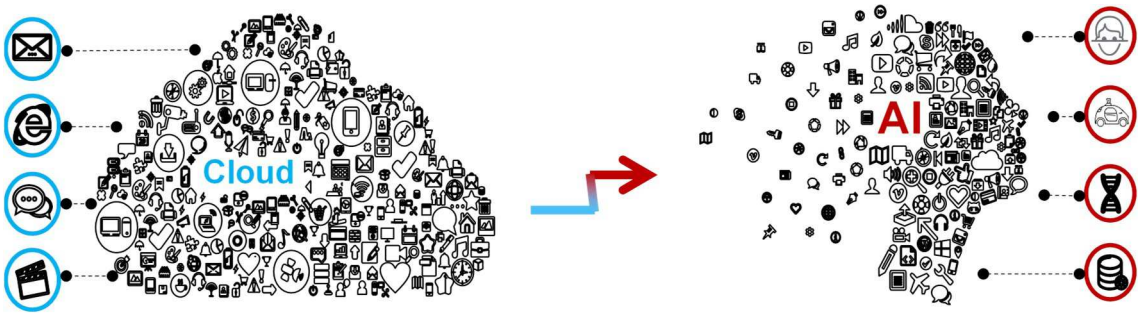


Figure 1—Digital transformation in the era of AI

Data centers in the cloud era focused on application transformation and the rapid deployment of services. During the AI era, data centers are the source of information and algorithms for the real-time digital transformation of our digital lives. The combination of high-speed storage and AI distributed computing render big data into fast data, access by humans, machines, and things. A high-performance, large scale data center network without packet loss is critical to the smooth operation of digital transformation.

For high-performance applications, such as AI, key measures of network performance include throughput, latency, and congestion. Throughput is dependent on the total capacity of the network to quickly transmit a large amount of data. Latency refers to the total delay for a transaction across the data center network. When the traffic load exceeds network capacity, congestion occurs. Packet loss is a factor that seriously affects both throughput and latency.

Currently, digital transformation of various industries is accelerating. It is estimated that 64% of enterprises have become the explorers and practitioners of digital transformation (Orange [38]). Among 2000 multinational companies, 67% of CEOs have made digitalization the core of their corporate strategies (Wiles [42]). The drive towards digital transformation in the real-time world is leading the data center network to support a 'Data-Centric' model of computing.

A large amount of data will be generated during the digitalization process, becoming a core asset, and enabling the emergence of artificial intelligence applications. The Huawei Global Industry Vision predicts that data volume will reach 180 ZB in 2025 [18]. However, data is not the "end-in-itself". Knowledge and wisdom extracted from data are eternal values. The proportion of unstructured data (such as raw voice, video, and image data) increases continuously, and will account for 95% of all data in the future (see Figure 2). Current big data analytic methods are unable to keep pace with the growth of data and performance innovations are needed to extract the value from the raw data. An AI approach based on deep learning can filter out massive amounts of invalid data and automatically extract useful information, providing more efficient decision-making and behavior guidance.

Copyright © 2021 IEEE. All rights reserved.
Documents Document Sharing Hub
Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 18,2021 at 02:08:55 UTC from IEEE Xplore. Restrictions apply.
Full document Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 18,2021

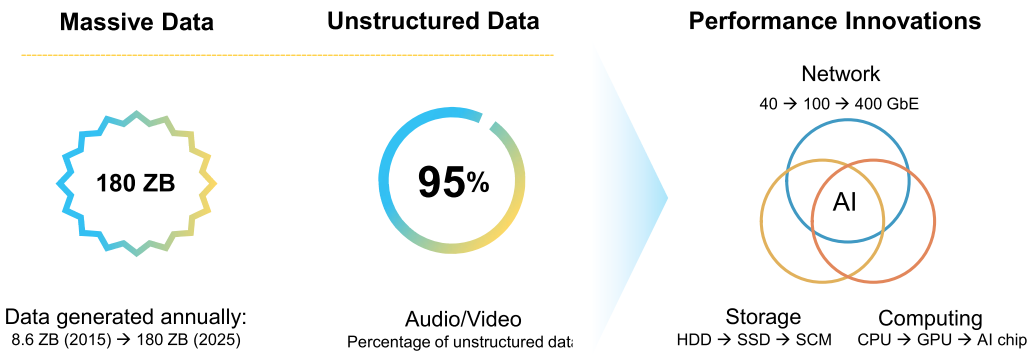


Figure 2—Emerging artificial intelligence applications

The cloud data center architecture improved the performance and scale of applications in general. The cloud platform allows rapid distribution of IT resources to create an application-centric service model. In the AI era, the applications are consuming unprecedented amounts of data and the cloud data center architecture is augmented with necessary performance innovations to handle the load. Seamlessly introducing these innovations along with new AI applications can be tricky in an existing cloud data center. Understanding how to efficiently process data based on the needs of AI applications is a key focus area. Orchestrating the flow of data between the storage and computing resources of the applications is a critical success factor.

3 Evolving data center requirements and technology

Previous Data Center Bridging Standards

During the early days of 10 Gbps Ethernet, the IEEE 802.1 Working Group developed a focus on Data Center Bridging (DCB). The DCB task group defined a set of enhancements to the Ethernet, Bridges, and associated protocols for use in data center environments. The use-case and focus were on clustering and storage area networks, where traditionally dedicated technologies such as InfiniBand™ and Fibre Channel were used.[2] Important objectives for Ethernet were to eliminate loss due to congestion and to allocate bandwidth on links for selected traffic. The key contributions at the time included the following:

- **Priority-based Flow Control (PFC)**: A link level flow control mechanism that eliminates packet loss and can be applied independently to each traffic class.
- **Enhanced Transmission Selection (ETS)**: A queue scheduling algorithm that allows for bandwidth assignments to traffic classes.

[2] InfiniBand is a trademark and service mark of the InfiniBand® Trade Association.

Copyright © 2021 IEEE. All rights reserved.
Documents Document Sharing Hub
Full document Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 18,2021 at 02:08:55 UTC from IEEE Xplore. Restrictions apply.

- **Congestion Notification**: A layer-2 end to end congestion management protocol that detects congestion, signals across the layer-2 network to limit the transmission rate of senders to avoid packet loss.
- **Data Center Bridging Capabilities Exchange Protocol (DCBX)**: A discovery and capability exchange protocol, working in conjunction with the Link Layer Discovery Protocol (LLDP), to convey capabilities and configuration of the above features.

These contributions were important to the expansion of Ethernet into the specialized markets of cluster computing and storage area networks. However, continued evolution is needed as the environments and technologies have changed. Today's data centers are deployed on massive scale, using Layer-3 protocols and highly orchestrated management systems. Ethernet links have advanced from 10 Gbps to 400 Gbps, with active plans to increase speeds into the Tbps range. New applications, such as Artificial Intelligence (AI) are placing new demands on the infrastructure and driving architectural changes. Continued innovation is needed to further expand the use of Ethernet in modern data center environments.

Requirements evolution

AI applications put pressure on the data center network. Consider AI training for self-driving cars as an example. The deep learning algorithm relies heavily on massive data and high-performance computing capabilities. The training data collected each day is approaching the petabyte level (1PB = 1024 TB), and if traditional hard disk storage and common CPUs were used to process the data, it could take at least one year to complete the training process. This is clearly impractical. To improve AI data processing efficiency, revolutionary changes are needed in the storage and computing fields. For example, storage performance needs to improve by an order of magnitude to achieve more than 1 million input/output operations per second (IOPS) (Handy and Coughlin [13]).

To meet real-time data access requirements, storage media has evolved from hard disk drives (HDDs) to solid-state drives (SSDs) to storage-class memory (SCMs). This has reduced storage medium latency by more than 1000 times. Without similar improvements in network latency, these storage improvements cannot be realized and simply move the bottleneck from the media to the network. With networked SSD drives, the communication latency accounts for more than 60% of the total storage end-to-end latency. With the move to SCM drives, this percentage could increase to 85% unless improvements in network performance are achieved. This creates a scenario where the precious storage media is idle more than half of the time. When considering recent improvements in both storage media and AI computing processors together, the communication latency accounts for more than 50% of the total latency, further hindering improvements and wasting resources (Huawei [16]).

AI applications and environments are growing in scale and complexity. For example, there were 7 ExaFLOPS and 60 million parameters in Microsoft's ResNet of 2015. Baidu used 20 ExaFLOPS and 300 million parameters when training their deep speech system in 2016. In 2017, the Google NMT used 105 ExaFLOPS and 8.7 billion parameters (Karuppiah [29]). New characteristics of AI computing are requiring an evolution of data center network.

Traditional protocols are no longer able to satisfy the requirements of new applications that serve our daily lives. In a simple example, the online food take-out industry at Meituan has increased nearly 500% in the last four years [43]. The number of transactions has increased from 2.149 billion to 12.36 billion where those transactions all occur within a few hours at peak mealtimes. The

Copyright © 2021 IEEE. All rights reserved.
Documents Document Sharing Hub
Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 18,2021 at 02:08:55 UTC from IEEE Xplore. Restrictions apply.
Full document Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 18,2021

Meituan Intelligent Scheduling System is responsible for orchestrating a complex multi-person, multi-point real-time decision-making process for end-users, businesses and over 600,000 delivery drivers. The drivers report positioning data 5 billion times a day that are used to calculate optional paths for the drivers and deliver optimal solutions within 0.55 milliseconds. When the back-end servers use TCP/IP protocols, the amount of data copied between kernel buffers, application buffers, and NIC buffers stresses the CPU and memory bus resources causing increased delay and an inability to meet the application requirements. The newer Remote Direct Memory Access (RDMA) protocol eliminates data copies and frees CPU resources to perform driver path and take-out order calculations at scale. The improved efficiency of RDMA puts more pressure on the network, moving the bottleneck to the data center network infrastructure where low-latency and lossless behavior become the new critical requirements.

Characteristics of AI computing

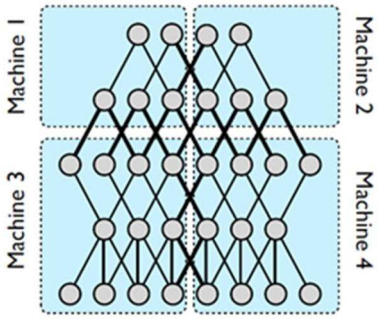
Traditional data center services (web, database, and file storage) are transaction-based and the calculated results are often deterministic. For such tasks, there is little correlation or dependency between a single transaction and the associated network communication. The occurrence and duration of the traditional transactions are random. AI computing, however, is different. It is an optimization problem with iterative convergence. This causes high spatial correlation within the data sets and computing algorithms and creates temporal correlations between communication flows.

AI computing works on big data and demands fast data. To achieve this, it must operate in parallel to “divide-and-conquer” the problem. The computing model and input data sets are large (e.g., in a 100 MB node, the AI model with 10K rules requires more than 4 TB memory). A single server cannot provide enough storage capacity and processing resources to handle the problem sequentially. Concurrent AI computing and storage nodes are required to shorten the processing time. This distributed AI computing and storage requirement highlights the need for a fast, efficient, and lossless data center network that has the flexibility to support two distinct parallel modes of operation—model parallel computing and data parallel computing.

Model parallel computing

In model parallel computing, each node computes one part of the overall algorithm. Each node processes the same set of data, but with a different portion of the algorithm, resulting in an estimate for a differing set of parameters. The nodes exchange their estimates to converge upon the best estimate for all the data parameters. With model parallel computing, there is an initial distribution of the common data set to a distributed number of nodes, followed by a collection of individual parameters from each of the participating nodes. Figure 3 shows how parameters of the overall model may be distributed across computing nodes in a model parallel mode of operation.

Copyright © 2021 IEEE. All rights reserved.
Documents Document Sharing Hub
Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 18,2021 at 02:08:55 UTC from IEEE Xplore. Restrictions apply.
Full document Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 18,2021

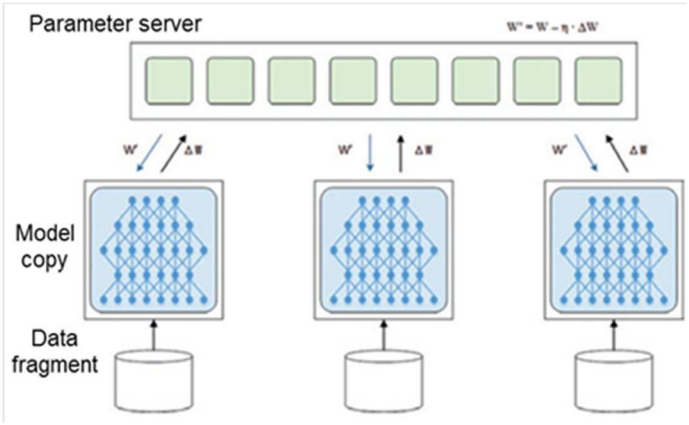


Reprinted with permission from the authors of "Large Scale Distributed Deep Networks" [9]

Figure 3—Model parallel training

Data parallel computing

In data parallel computing, each node loads the entire AI algorithm model, but only processes part of the input data. Each node is trying to estimate the same set of parameters using a different view of the data. When a node completes a round of calculations, the parameters are weighted and aggregated by a common parameter server as seen in Figure 4. The weighted parameter update requires that all nodes upload the information synchronously.



Reprinted with permission from the authors of "Large Scale Distributed Deep Networks" [9]

Figure 4—Data parallel training

Regardless of the parallel computing approach, data center networks feel the pressure of demanding communication. When the network becomes the bottleneck, the waiting time for computing resources can exceed 50% of the job completion time (Cardona [4]).

Copyright © 2021 IEEE. All rights reserved.
Documents Document Sharing Hub
Full document Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 18,2021 at 02:08:55 UTC from IEEE Xplore. Restrictions apply.

With all AI applications, the computing model is iterative and requires a synchronization step that creates network incast congestion. Figure 5 shows how incast congestion occurs with AI training. The training process is iterative and there are many parameters synchronized on each iteration. The workers download the model and upload newly calculated results (ΔM) to a parameter server during a synchronization step. The uploading to the parameter server creates incast. When the computing time is improved by deploying new compute technology, the pressure on the network and resulting incast increases.

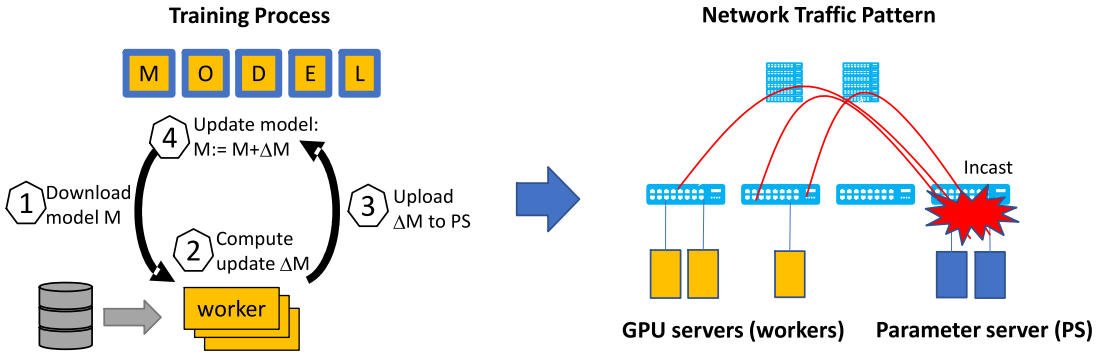


Figure 5—Periodic incast congestion during training

The communication between the worker nodes and the parameter server constitutes a collection of interdependent network flows. In the iteration process of distributed AI computing, many burst traffic flows are generated to distributed data to workers within milliseconds, followed by an incast event of smaller sized flows directed at the parameter server when the intermediate parameters are delivered and updated. During the exchange of these flows, packet loss, congestion, and load imbalance can occur on the network. As a result, the Flow Completion Time (FCT) of some of the flows is prolonged. If a few flows are delayed, storage and computing resource can be underutilized. Consequently, the completion time of the entire application is delayed.

Distributed AI computing is synchronous, and it is desirable for the jobs to have a predictable completion time. When there is no congestion, dynamic latency across the network is small allowing the average FCT to be predictable and therefore the performance of the entire application is predictable. When congestion causes dynamic latency to increase to the point of causing packet loss, FCT can be very unpredictable. Flows that complete in a time that is much greater than the average completion contributes to what is known as tail latency. Tail latency is the small percentage of response times from a system, out of all of responses to the input/output (I/O) requests it serves, that take the longest in comparison to the bulk of its response times. Reducing tail latency as much as possible is extremely critical to the success of parallel algorithms and the whole distributed computing system. To maximize the use of computing resources in the data center, tail latency should be addressed.

Evolving technologies

Progress can be seen when evolving requirements and evolving technologies harmonize. New requirements often drive the development of new technologies and new technologies often enable new use cases that lead to, yet again, a new set of requirements. Breakthroughs in networked

Copyright © 2021 IEEE. All rights reserved.
Documents
Document Sharing Hub
Full document Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 18,2021 at 02:08:55 UTC from IEEE Xplore. Restrictions apply.

storage, distributed computing, system architecture, and network protocols are enabling the advancement of the next generation data center.

SSDs and NVMeoF: High throughput, low-latency network

In networked storage, a file is distributed to multiple storage servers for I/O acceleration and redundancy. When a data center application reads a file, it accesses different parts of data from different servers concurrently. The data is aggregated through a data center switch at nearly the same time. When a data center application writes a file, the writing of data can trigger a series of storage transactions between distributed and redundant storage nodes. The entire sequence of transactions must complete before the original write action is satisfied. Figure 6 shows an example of data center communication triggered by the networked storage service model.

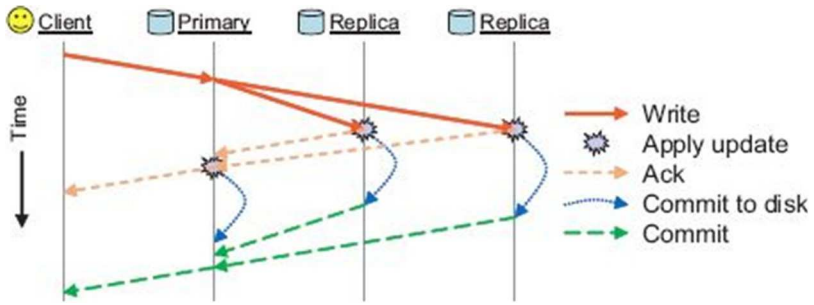


Figure 6—Networked storage service model

The example highlights the importance of the network enabling both high throughput and low latency simultaneously. The bulk data being written to the primary storage server is transmitted multiple times to the replicas. The small sized acknowledgments and commit messages must be sequenced and ultimately delivered to the originating client before the transaction can complete, emphasizing the need for ultra-low latency.

Massive improvements in storage performance have been achieved as the technology has evolved from HDD to SSD to SCM using the Non-Volatile Memory Express (NVMe) interface specification. Accessing storage media via NVMe has decreased access time by a factor of 1000 over previous HDD technology. Sample seek times between the various technologies include: HDD = 2-5 ms, SATA SSD = 0.2 ms, and NVMe SSD = 0.02 ms. SCM is generally three to five times faster than NVMe flash SSDs.

NVMe-over-fabrics (NVMeoF) involves deploying NVMe for networked storage. The much faster access speed of the medium results in greater network bottlenecks and the impact of network latency becomes more significant. Figure 7 shows how network latency has become the primary bottleneck with faster NVMe based storage. Network latency was a negligible part of end-to-end networked HDD storage latency but is poised to become a significant component of latency with networked SCM storage. To maximize the IOPS performance of the new medium, the network latency problem needs to be resolved first.

Copyright © 2021 IEEE. All rights reserved.
Documents Document Sharing Hub
Full document Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 18,2021 at 02:08:55 UTC from IEEE Xplore. Restrictions apply.

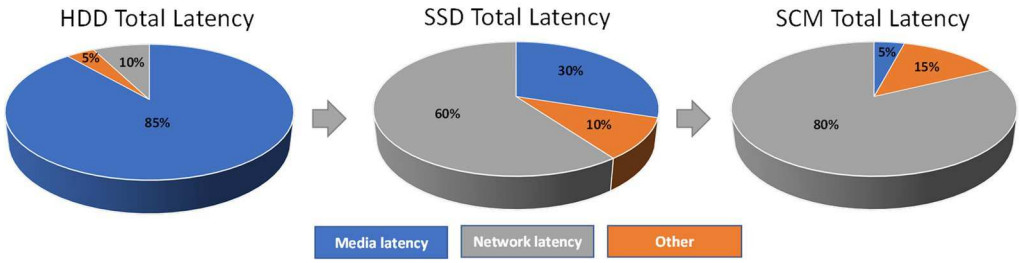


Figure 7—End-to-end latency breakdown for HDD and SSD

There are two distinct types of latency—static latency and dynamic latency. Static latency includes serial data latency, device forwarding latency, and optical/electrical transmission latency. This type of latency is determined by the capability of the switching hardware and the transmission distance of the data. It usually is fixed and very predictable. Figure 8 shows that current industry measurements for static latency are generally at the nanosecond (10^{-9} second) or sub-microsecond (10^{-6}) level, and account for less than 1% of the total end-to-end network delay.

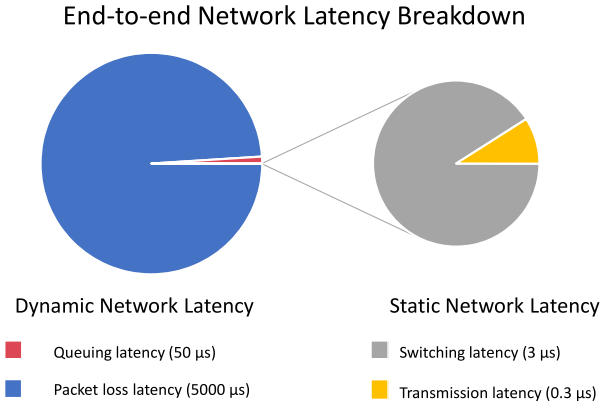


Figure 8—Network latency breakdown

Dynamic latency plays a much greater role in total end-to-end network delay and is greatly affected by the conditions within the communication environment. Dynamic latency is created from delays introduced by internal queuing and packet retransmission, which are caused by network congestion and packet loss. Parallel AI computing models create unique traffic patterns that result in heavy network congestion. The key to low end-to-end network latency is to address dynamic latency and the key to addressing dynamic latency is mitigating congestion.

The major component of dynamic latency is the delay from packet retransmission when packets are dropped within the network. Packet loss latency is an order magnitude greater than queuing delay and has proven to have a severe impact on applications. Packet loss occurs when switch buffers are overrun because of congestion. (Please note that we ignore packet loss due to low-probability bit errors during transmission.) There are two key types of congestion that lead to packet loss—in-network congestion and incast congestion. In-network congestion occurs on switch-to-switch links within the network fabric when the links become overloaded, perhaps due to ineffective load balancing. Incast congestion occurs at the edge of the network when many sources are sending to

Copyright © 2021 IEEE. All rights reserved.
Documents Document Sharing Hub
Full document Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 18,2021 at 02:08:55 UTC from IEEE Xplore. Restrictions apply.