Style Definition: Table Text

IEEE 802 Nendica Report: Intelligent Lossless Data Center Networks

Editor

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

Nendica Chair

Name	Affiliation
Marks, Roger	Huawei

Contributors/Supporters

Name	Affiliation
Li, Jie	CIACT/ODCC
Gao, Feng	Baidu
Gu, Rong	China Mobile
Zhao, Jizhuang	China Telecom
Chen, Chuansheng	Tencent
Yin, Yue	Huawei
Song, Qingchun	Mellanox
Lui, Jun	Cisco
He, Zongying	Broadcom
Sun, Liyang	Huawei

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.

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

PDF: ISBN xxx-x-xxxx-xxxx-x XXXXXXXXXX

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

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.

TABLE OF CONTENTS

1 Introduction

<<Editor's notes will be noted inside these marking and removed in future drafts>>

<<short intro and the more detailed background intro is section 2. This will be written near the end>>

This paper is the result of the Data Center Networks work item [1] within the IEEE 802 "Network Enhancements for the Next Decade" Industry Connections Activity known as Nendica. The paper is an update to a previous report, IEEE 802 Nendica Report: The Lossless Network for Data Centers published on August 17, 2018 [2]. This update provides additional background on evolving use cases in modern data centers and proposes solutions to new problems identified by this paper.

Scope

The scope of this report includes...

Purpose

The purpose of this report is to ...

2 Bringing the data center to life

A new world with data everywhere

<<

- ✓ Enterprise digital transformation needs more data for using AI
- ✓ Machine translation and search engines need to be able to process huge data simultaneously
- ✓ The era of internet celebrity webcast, all-people online games, data explosion
- ✓ Consumption upgrade in the new era of take-out, online takeout platform schedule and deliver massive orders
- ✓ The XX service of the carrier has higher requirements on data center network
- ✓ Data-based New World Requires Ubiquitous Data Center Technologies.

>>

Digital transformation is driving change in both our personal and professional lives. Work flows 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 timely information, automated human interactions, and refined decision making. 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 immediate results. Data center networks must deliver unprecedented levels of performance and reliability to meet these real-time demands.

For high-performance applications, such as AI, key measures for network performance include throughput, latency, and congestion. Throughput is dependent on the total capacity of the network for quickly transmitting a large amount of data. Latency refers to the total delay on the network when performing a transaction across the data center network. When the traffic load exceeds the network capacity, congestion occurs. Packet loss is a factor that seriously affects both throughput and latency. Data loss in a network may cause a series events that deteriorate performance. For example, an upper-layer application may need to retransmit lost data in order to continue. Retransmissions can increase load on the network, causing further packet loss. In some applications, delayed results are not useful, and the ultimate results can be discarded, thus wasting resources. In other cases, the delayed result is just a small piece of the puzzle being assembled by the upper-layer application that has now been slowed down to the speed of the slowest worker. More seriously, when an application program does not support packet loss and cannot be restored to continue, a complete failure or damage can be caused.

Data centers ultimately deliver the services in this era of digital transformation to our real-time 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 the modern data center.

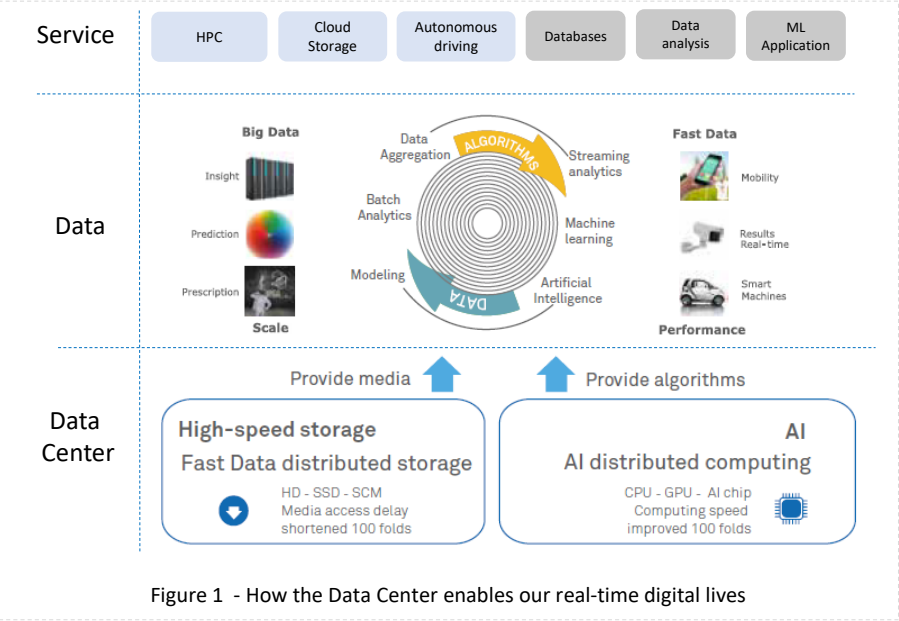


Figure 1 - How the Data Center enables our real-time digital lives

Today's data center enables the digital real-time world

Currently, digital transformation of various industries is accelerating. According to analysis data, 64% of enterprises have become the explorers and practitioners of digital transformation <<IDC reference>>. Among 2000 multinational companies, 67% of CEOs have made digitalization the core of their corporate strategies ~~<<Gartner reference>>.~~[3].

A large amount of data will be generated during the digitalization process, becoming a core asset, and enabling a new emergence of Artificial Intelligence Applications as seen in Figure W. Huawei GIV predicts that the data volume will reach 180 ZB in 2025 ~~<<Huawei reference>>.~~[4]. However, data is not the "end-in-itself". Knowledge and wisdom extracted from data are eternal values. However, the proportion of unstructured data (such as raw voice, video, and image data) increases continuously, and will reach over 95% in the future. The current big data analytics method is helpless. If manual processing is used, the data volume will be far greater than the processing capability of all human beings. The AI algorithm based on machine computing for deep learning can filter out massive invalid data and automatically reorganize useful information, providing more efficient decision-making suggestions and smarter behavior guidance.

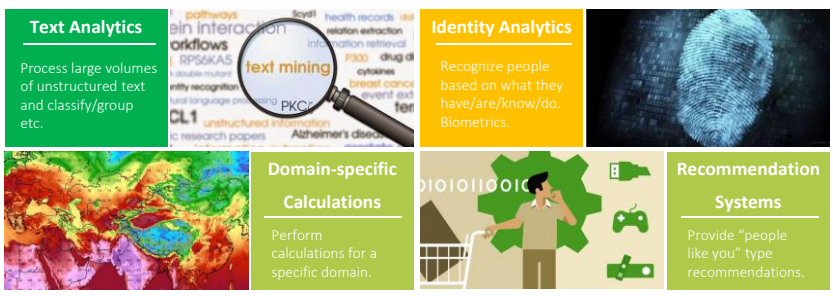


Figure 2 – Emerging Artificial Intelligence Applications

AI applications are emerging everywhere, as shown in Figure 2.

Cloud data centers improve the performance of these applications. Cloud data centers are designed to scale and act more like a service support center. They are application-centric and use the cloud platform to quickly distribute IT resources. While the data centers are application centric, they are founded on big data as shown in Figure 3.

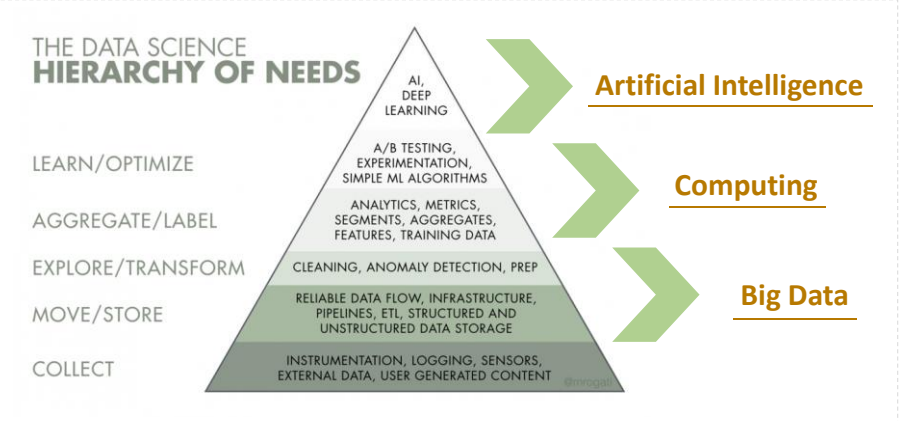


Figure 3 – Data is the Foundation for Artificial Intelligence

So, within data centers, understanding how to efficiently process data based on the needs of different applications is a key focus area. Data centers must know where to reserve storage to efficiently transmit the data to the computing engines of the applications.

~~<< Notes:~~

- ~~**AI related services**: AI cloud data center improves these applications performance: smart manufacturing/finance/energy/transportation (cloud data centers go to AI era. A cloud data center is more like a service support center. It is application-centric and uses the cloud platform to quickly distribute IT resources. The data center for AI services evolves into a business value center based on the cloud data center. The data center focuses on how to efficiently process data based on AI)~~
- ~~**Distributed storage**: Stay ahead of rapid storage growth driven by new data sources and evolving technologies, a flexible storage efficiency is critical for customers to maximize the revenue of every bit. The development of high speed storage technology will help users to access the content more conveniently. Other data center technologies should be evolved together with distributed storage to ensure customers can obtain high input and output speed.~~
- ~~**Cloud Database**: A cloud database may be a native service within a public cloud provider, or it may be a database from a cloud agnostic software vendor, designed for cloud architectures and requirements. Data centers make use of new technologies to address distributed cloud databases modern high performance application requirements.~~

~~>>~~

3 Evolving data center requirements and technology

Requirements evolution

<< First discuss the new and evolving requirements for data center networks hosting AI applications. These include: 1. huge amounts of data for AI learning. What is the data and why is it so large? Why is AI better with more data? 2. To hold the huge amounts of data, it must be fast and distributed. The latency needs to compete with local storage. 3. Huge amounts of computing cycles needed to work on that data. Describe the AI computing models; data parallelism vs model parallelism and how these differ in network communication requirements. >>

Take AI training of self-driving cars as an example, the deep learning algorithm relies heavily on massive sample data and high-performance computing capabilities. Training data collected is close to the P level (1PB = 1024 TB) per day. If traditional hard disk storage and common CPUs are used to process the data, it takes at least one year to complete the training, which is almost impossible. To improve AI data processing efficiency, revolutionary changes are occurring in the storage and computing fields. The development of high-speed storage technology will help users to access the content more conveniently. Other data center technologies should be evolved together with distributed storage to ensure customers can obtain high input and output speed. Storage performance needs to improve by an order of magnitude to achieve more than 1 million input/output operations per second (IOPS) [3].

Storage media evolve from HDDs to SSDs to meet real-time data access requirements, reducing the medium latency by more than 100 times. With the significant improvement of storage media and computing capabilities, the current network communication latency becomes the bottleneck of further performance improvement in high-performance data center clusters. The communication latency accounts for more than 60% of the total storage E2E latency, that is, more than half of the time of precious storage media is idle.

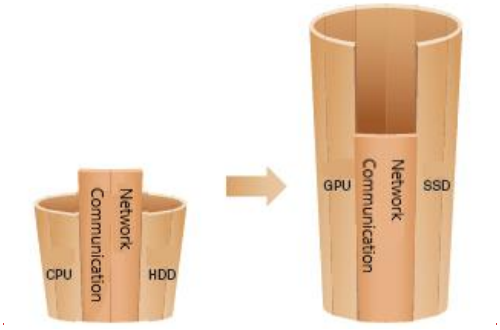


Figure Z – Network performance becomes the shortest plank

In general, with the evolution of storage media and computing processors, the communication duration accounts for more than 50% of the total communication duration, hindering the further improvement of computing and storage efficiency [4]. ~~Only when the communication duration is reduced close to time cost of computing and storage, the 'short planks' in the bucket principle can be eliminated (see in Figure Z), and the computing and storage performance can be effectively improved.~~

- ✓ The development of fast storage provides necessary media for big data (distributed storage)
  - Storage performance needs to improve by an order of magnitude to achieve more than 1 million input/output operations per second (IOPS).
  - Communication latency has recently increased from 10% to 60% of storage E2E latency.
- ✓ Computing speed improvement (distributed computing)

AI computing model complexity is exploding

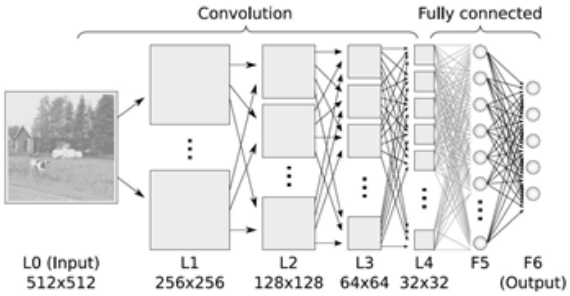


Figure A - Iterative machine learning network model

AI training is becoming increasingly complex with the development of services. For example, there are 7 ExaFLOPS and 60 million parameters in the Microsoft Resnet in 2015. The number came to 20 ExaFLOPS and 300 million parameters when Baidu trained their deep speech system in 2016. In 2017, the Google NMT used 105 ExaFLOPS and 8.7 billion parameters [5].

AI inference is the next great challenge so there must be an explosion of network design. The new characteristics of AI algorithm and huge computing workload require evolution of data center network.

Characteristics of AI computing

<< explain that AI computing is iterative, not a single pass, so communication is critical and the application runs for a long time. Describe the different modes for AI computing; data parallel vs model parallel and what that means to the network >>

Traditional data center services (web, video, and file storage) are transaction-based and the calculation results are deterministic. For such tasks, there is no correlation or dependency between single calculation and network communication, and the occurrence time and duration of the entire calculation and communication are random. AI computing is based on target optimization and iterative convergence is required in the computing process, which causes high spatial correlation in the computing process of AI services and temporally similar communication modes.

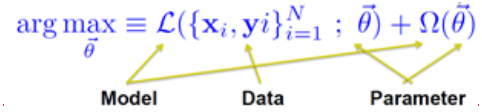


Figure B - Machine learning algorithm

A typical AI algorithm refers to an optimization process for a target. The computing scale and features mainly involve models, input data, and weight parameters.

To solve the Big Data problem, the computing model and input data need to be large (for a 100 MB node, the AI model for 10K rules requires more than 4 TB memory), for which a single server cannot provide enough storage capacity. In addition, because the computing time needs to be shortened and increasingly concurrent AI computing of multiple nodes is required, DCNs must be used to perform large-scale and concurrent distributed AI computing.

Distributed AI computing has the following two modes: model parallel computing and data parallel computing. For model parallel computing, each node computes one part of the algorithm. After computing is complete, all data fragmented across models needs to be transferred to other nodes, as shown in Figure C.

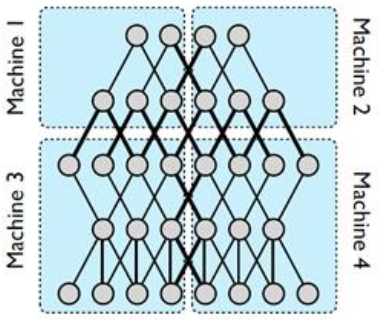


Figure C - Model parallel training

For parallel data computing, each node loads the entire AI algorithm model. Multiple nodes can calculate the same model at the same time, but only part of the input data is input to each node. When a node completes a round of calculation, all relevant nodes need to aggregate updated information about obtained weight parameters, and then obtain the corresponding globally updated data. Each weight parameter update requires that all nodes upload and obtain the information synchronously.

Commented [PC1]: A good summary of the differences is:

'Data parallelism' and 'model parallelism' are different ways of distributing an algorithm. These are often used in the context of machine learning algorithms that use stochastic gradient descent to learn some model parameters, which basically means that:

- The algorithm is trying to estimate some parameters from the given data.
- Parameters are estimated by minimizing the gradient against some loss function.
- Algorithm iterates over data in small batches.

In the data-parallel approach:

- The algorithm distributes the data between various cores.
- Each core independently tries to estimate the same parameter(s)
- Cores then exchange their estimate(s) with each other to come up with the right estimate for the step.

In the model-parallel approach:

- The algorithm sends the same data to all the cores.
- Each core is responsible for estimating different parameter(s)
- Cores then exchange their estimate(s) with each other to come up with the right estimate for all the parameters.

Data-parallel approach is useful when there are smaller number of nodes in the cluster and the number of parameters to be estimated is small whereas model-parallel approach is useful in the opposite condition.

No matter the development of distributed storage or distributed AI training, data center network comes to the communication pressure. The waiting time for GPU communication exceeds 50% of the job completion time [6].

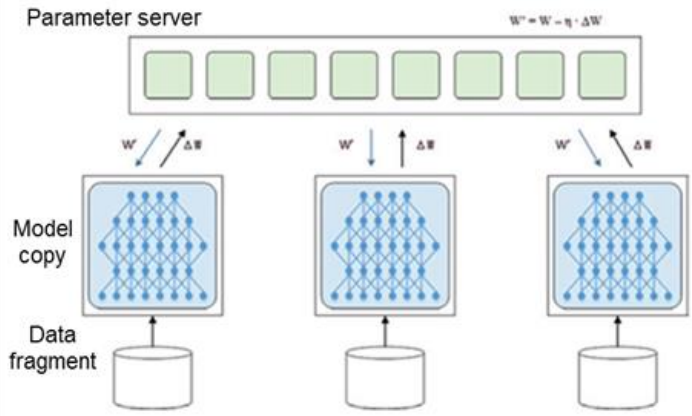


Figure D - Data parallel training

Evolving technologies

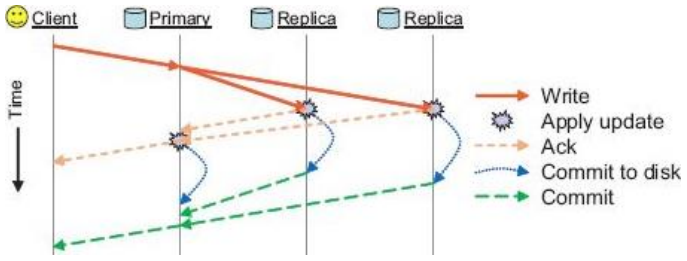
<< Here we describe some key new technologies that are evolving to meet the requirements. The main pieces should be fast storage (SSDs), GPUs, Smart Nics, Protocols like RDMA. Have a small sub-section on each of the above technology areas. End with what it means to the network to support these new technologies >>

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

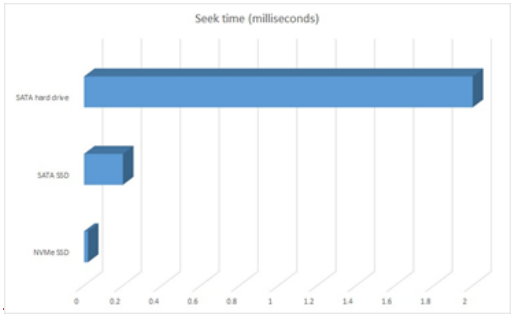
SSDs and NVMeoF: High throughput, low-latency network

In networked storage, a file is distributed to multiple storage servers for IO acceleration and redundancy. When a data center application reads a file, it will concurrently access different parts of data from different servers, and the data will be aggregated through a data center switch at nearly the same time. When a data center application writes a file, the data can trigger a series of storage transactions between distributed and redundant storage nodes. Figure 8 shows an example of data center communication triggered by the networked storage service model.

When an application (i.e. Client in Figure 8) requests to write a file, it will concurrently send data to the object storage device (OSD) servers. There are two types of OSD servers, one type is the primary, and the other type is the replica. When the primary servers receive data that need to be saved, it will transmit the data to the replica servers twice as backup (the orange arrowhead in Figure 8). After receiving the data, the primary OSD server will send an ACK to client while the replica servers will send ACK to the primary server (pink dash line in Figure 8). Each OSD server will then begin to commit the data to the storage medium. It takes a short period time to commit and store data. When the replica servers finish saving data, they will send commit notification to primary server to notify that the writing task is complete. Once the primary server has received all the commit information from all replica servers, the primary server will send a commit message to client. The storage write process is not complete until the primary server has sent the final commit message to the client. << Consider making a comment about the impact of network latency here >>.



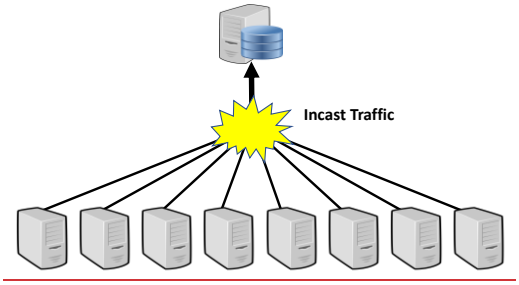
~~Figure 8 - Networked storage service model~~



~~Figure 9 - Performance improvement of storage media~~

Massive improvements in storage performance have been achieved as the technology has evolved from HDD to SDD to NVMe (Non-Volatile Memory Express). The latest storage media technology, NVMe, has decreased access time by a factor of 1000 over previous HDD technology. Figure 9 shows the difference in seek time between the technologies; HDD = 2-5 ms, SATA SSD = 0.2 ms, and NVMe SSD = 0.02 ms. Shorter overall average seek times are better, but performance of drives in each category can vary [8].

When NVMe is used for networked storage, the much faster access speed of the medium can result in network bottlenecks. Figure 10 shows a classical networked storage traffic model. In this traffic



~~Figure 10 - Distributed storage traffic characteristics~~

model, when data is aggregated each time, incast (many-to-one) easily to occurs. As storage performance continues to increase, pressure on the network increases, affecting distributed storage IO throughput.

~~<< NOTE: the figure 10 doesn't really make sense with the discussion. We talked about distributed (networked) storage where data is in multiple physical locations and an application reads/writes causing network traffic. Here, we show a bunch of servers writing to a single disc.>>~~

Incast is a network traffic pathology that affects many-to-one communication patterns in datacenters. Incast increases application latency with the queuing delay of flows and decreases application throughput to something well below the link bandwidth [9]. The problem especially affects computing paradigms, such as AI training, where distributed processing cannot continue until all parallel threads in a stage complete.

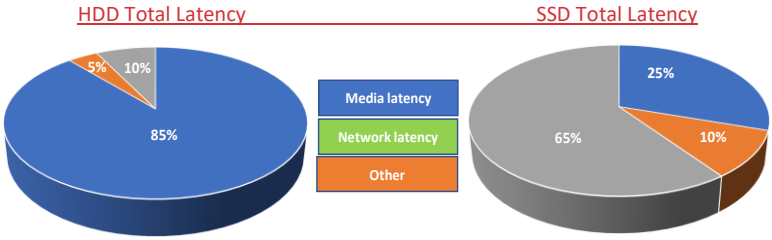


Figure 11 – End-to-end latency breakdown for HDD and SDD scenarios

Since incast increases application latency, the concurrency of the networked storage system will be affected. Therefore, the performance of distributed IOPS is limited by network latency. With newer, faster storage technologies, the impact of network latency becomes more significant. Figure 11 shows that network latency is the primary bottleneck in networked SSD storage, whereas network latency was negligible with networked HDD storage. Looking to the future, with NVMe over fabrics (i.e. networked NVMe storage), to attain the maximum IOPS performance, the network latency problem must be resolved first.

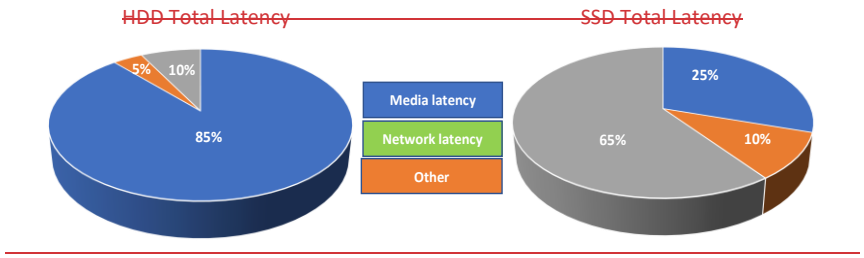


Figure 11 – End-to-end latency breakdown for HDD and SDD scenarios

<<

- High I/O throughput with low-latency storage network
  - As media access speeds increase, network latency becomes the bottleneck
  - Storage interface protocols evolve from Serial Attached SCSI (SAS) to Non-Volatile Memory Express (NVMe)
  - Reducing dynamic latency (latency from queuing and packet loss) is key to reducing the NVMe over Fabric latency

To analysis network latency further, it can be classified into 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 forwarding chip and the transmission distance. It usually has a fixed specification. Figure X says that static latency is generally at ns (10-9 second) or sub-us (10-6) level in the industry, and accounts for less than 1% of the total network delay.

The dynamic latency greatly affects the network performance. The dynamic latency ratio is greater than 99%. The dynamic latency includes the internal queuing latency and packet retransmission latency, which are caused by network congestion and packet loss. In the AI era, traffic conflicts become more and more severe on networks. Packet queuing or packet loss often occurs, causing the latency within sub-seconds. Therefore, the key of the low-latency network is the low dynamic latency.

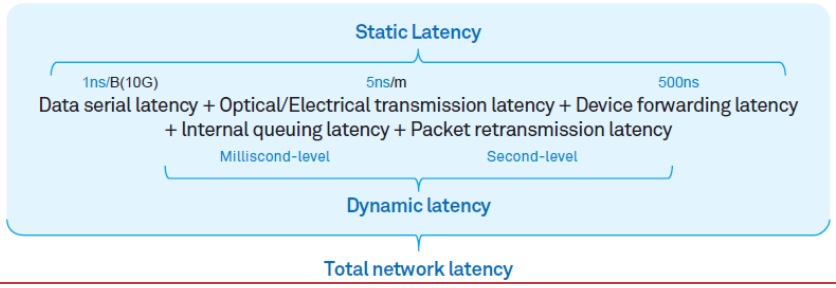


Figure XX – Total Network Latency

Most existing network solutions focus on reducing the static latency caused by network device forwarding, while the dynamic latency caused by packet loss during network congestion has proven to have a more severe impact on applications. In most existing systems the impact from latency comes primarily from dynamic latency which occurs across the network during packet loss of congestion management. Figure X below shows a typical network latency distribution.

<< NOTE: We should define dynamic latency and its components. Highlight the issue of packet loss. That has not been discussed yet. >>

Formatted: List Paragraph

Formatted: Not Highlight

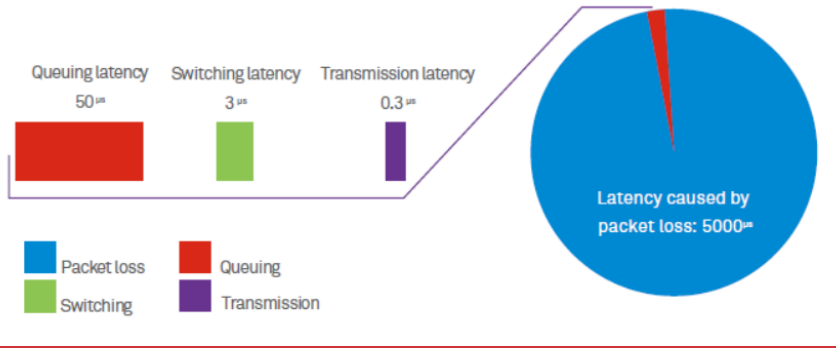


Figure YY – Latency Breakdown

>>

GPUs: Ultra-low latency network for parallel computing

As the number of AI algorithms and AI applications continue to increase, and the distributed AI computing architecture emerges~~, AI computing is implemented on a large scale.~~ , AI computing is implemented on a large scale. GPUs have ignited a worldwide AI boom. They have become a key part of modern supercomputing. They've been woven into a sprawling new hyperscale data centers. Still prized by gamers, they have become accelerators speeding up all sorts of tasks from encryption

Commented [PC2]: What is this architecture? Perhaps we should spend some time describing what it is and how the GPU plays a role. We don't provide any background on GPUs yet

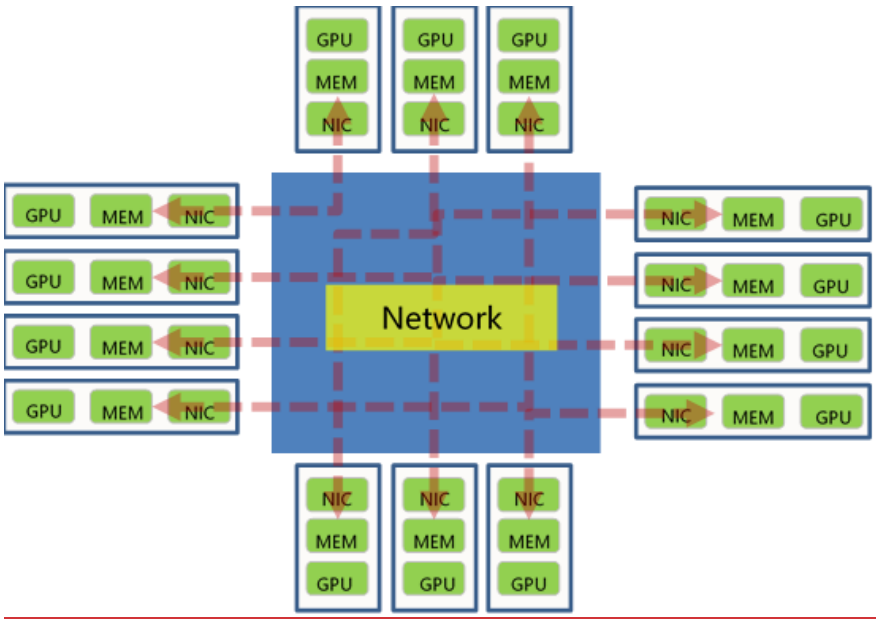


Figure WW – Distributed AI Computing Architecture

to networking to AI. GPUs provide much higher memory bandwidth than today's CPU architectures. Nodes with multiple GPUs are now ubiquitous in high-performance computing because of their power efficiency and hardware parallelism. Figure X illustrates the architecture of typical multi-GPU nodes, each of which consists of a host (CPUs) and several GPU devices connected by a PCI-e switch or NVLink. Each GPU is able to directly access its local relatively large device memory, much smaller and faster shared memory, and a small pinned area of the host node's DRAM, called zero-copy memory [11].

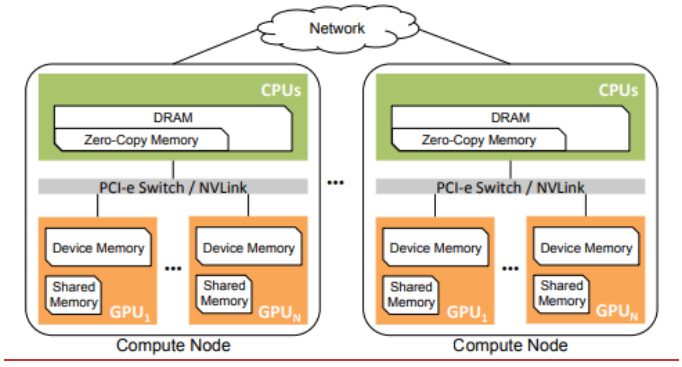


Figure WW2 – Distributed AI Computing Architecture

Today's AI computing architecture includes a hybrid mix of Central Processing Units (CPUs) and Graphics Processing Units (GPUs). GPUs, originally invented to help render video games at exceptional speeds, have found a new home in the data. The GPU is a processor with thousands of cores capable of performing millions of mathematical operations in parallel. All AI learning algorithms perform complex statistical computations and deal with a huge number of matrix multiplication operations per second – perfectly suited for a GPU. However, to scale the AI computing architecture to meet the needs of today's AI algorithms and applications in a data center, the GPUs must be distributed and networked. This places stringent requirements on communication volume and performance.

Facebook recently tested the distributed machine learning platform Caffe2, in which the latest multi-GPU servers are used for parallel acceleration. In the test, computing tasks on eight servers resulted in underutilized resources on the 100 Gbit/s InfiniBand network. The presence of the network and network contention reduced the performance of the solution to less than linear scale [10]. Consequently, network performance greatly restricts horizontal extension of the AI system.

GPUs are inherently designed to work on parallel problems. With AI applications, these problems are iterative and require a synchronization step that creates network incast congestion. Figure 12 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 parameter servers at nearly the same time. When the computing time is improved by deploying faster GPUs, the pressure on the network and resulting incast increases.

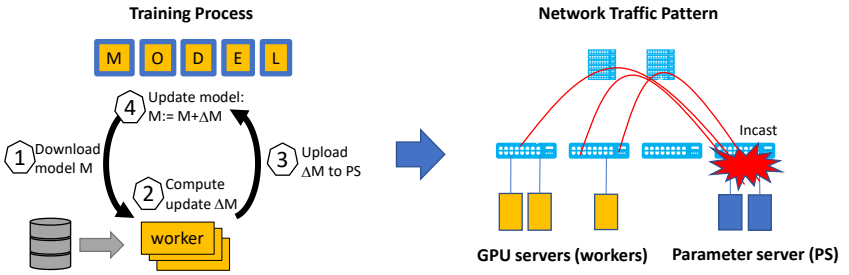


Figure 12 - Periodic incast congestion during training

The high-bandwidth and low-latency DCN with only physical links cannot meet requirements of large-scale and highly concurrent AI/HPC applications. In the iteration process of distributed AI computing, a large amount of burst traffic is generated within milliseconds. In addition, because a parameter server (PS) architecture is used to update parameter weights of the new model for data parallelization, the incast traffic model at a fixed time is easily formed. In this case, packet loss, congestion, and load imbalance occur on the network. As a result, the Flow Completion Time (FCT) of some data flows is too long. Distributed AI computing is synchronous. If few flows are delayed, more computing processes are affected. Consequently, the completion time of the entire application is delayed. This is what we call the 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. It is very critical to the whole distributed computing system. Figure X shows how tail latency injures the whole system performance.

Consequently, in order to minimize the FCT to complete the entire computing task, we need to reduce the tail delay as much as possible. Because the microbursts in data center network are within milliseconds, the tail delay needs to be controlled within milliseconds to ensure optimal system performance. Therefore, For HPC services, in order to have an ultra-low latency lossless network, the data center network should first solve the tail delay problem.

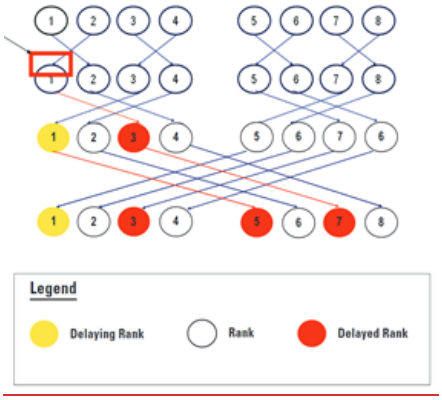


Figure X - Single-point delay causes the overall performance to deteriorate

~~✓ Ultra-low latency network for distributed computing~~

~~DCN Requirement of distributed AI computing~~

~~As the number of AI algorithms and AI applications continue to increase, and the distributed AI computing architecture emerges, AI computing has become implemented on a large scale. To ensure enough interaction takes place between such distributed information, there are more stringent requirements regarding communication volume and performance. Facebook recently tested the distributed machine learning platform Caffe2, in which the latest multi-GPU servers are used for parallel acceleration. In the test, computing tasks on eight servers resulted in insufficient resources on the 100 Gbit/s InfiniBand network. As a result, it proved difficult to achieve linear computing acceleration of multiple nodes. The network performance greatly restricts horizontal extension of the AI system.~~

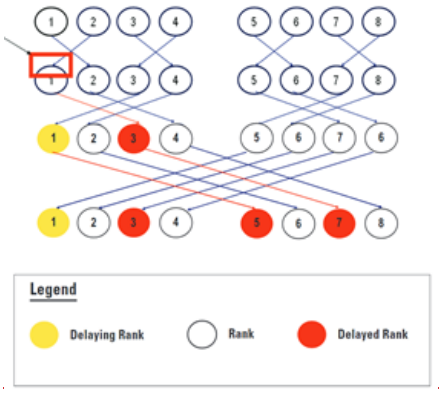
- ~~Controlling the tail latency of these applications is critical. It must be measured in microseconds, not milliseconds~~

SmartNICs

✓ SmartNIC become the computer in front of computer

- SmartNIC is a NIC with all NIC functions regardless CPU/FPGA. Host CPU only request to install NIC driver.

- SmartNIC is a computer in front of computer. SmartNIC has independent OS and is able to run some applications independently.
  - SmartNIC can be used to accelerate application
    - Accelerate computing, storage…
  - SmartNIC can be used to offload host CPU to run specific application more efficient
  - SmartNIC is part of computing resource. Participate the application computing together with host CPU and GPU.
    - Complement of CPU and GPU computing resource
    - SmartNIC is not the replacement of CPU and GPU, major applications still run on CPU/GPU
  - SmartNIC can be the independent domain than host domain and protect the host domain
    - Offload OVS to SmartNIC to isolate the data classification from hypervisor
  - SmartNIC can be emulated to other PCIe devices to support more advanced application
    - NVMe emulation
- SmartNIC is programmable and easy use
  - Open source software, major Linux
  - Easy to program, no special request for programmer
- SmartNIC is not proprietary NIC, one NIC fits many applications, easy for user to program



~~Figure X - Single-point delay causes the overall performance to deteriorate~~

RDMA

RDMA (Remote Direct Memory Access) is a new technology designed to solve the problem of server-side data processing latency in network applications, which transfers data directly from one computer's memory to another without the intervention of both operating systems. This allows for high bandwidth, low latency network communication and is particularly suitable for use in massively parallel computer environments. By transferring telegrams directly into the storage space of the other computer through the network, data can be quickly transferred from one system to the storage space of another system, reducing or eliminating the need for multiple copies of data telegrams during transmission, thus freeing up memory bandwidth and CPU cycles and greatly improving system performance. Figure E shows the principle of RDMA protocol.

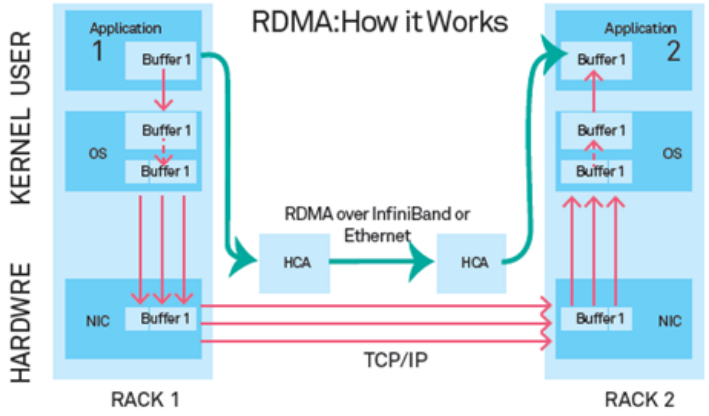


Figure E - Working principle of RDMA

RDMA's development in the transport layer/network layer currently goes through 3 technologies, Infiniband, iWarp and RoCEv1/RoCEv2.

Infiniband

In 2000, the IBTA (InfiniBand Trade Association) released the first RDMA technology, Infiniband, which is a customized network technology for RDMA multi-layered, new design from the hardware perspective to ensure the reliability of data transmission. The InfiniBand technology uses RDMA technology to provide direct read and write access to remote nodes. RDMA used InfiniBand as the transport layer in its early days, so it must use InfiniBan switches and InfiniBan network cards to implement.

iWarp

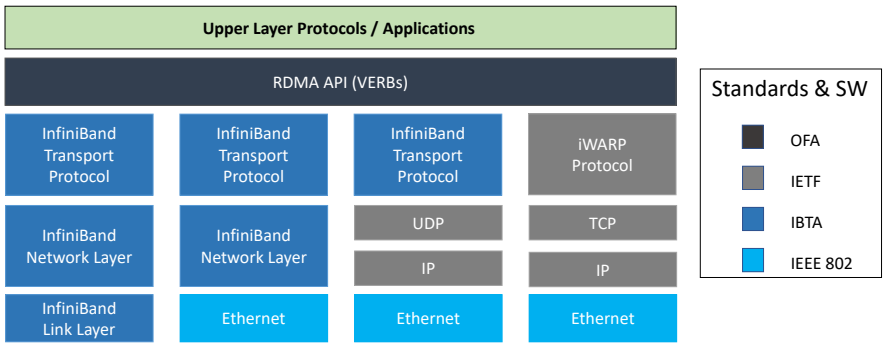


Figure xx – RDMA protocol stacks and standards – NOTE: SNIA Copyright

Internet wide area RDMA protocol, also known as RDMA over TCP protocol, is the IEEE/IETF proposed RDMA technology. It uses the TCP protocol to host the RDMA protocol. This allows RDMA to be used in a standard Ethernet environment (switch) and the network card requirement is an iWARP enabled network card. In fact iWARP can be implemented in software, but this takes away the performance advantage of RDMA.

RoCE (RDMA over Converged Ethernet)

In April 2010, the IBTA released RoCEv1, which was released as an add-on to the Infiniband Architecture Specification, so it is also known as IBoE (InfiniBand over Ethernet).The RoCE standard replaces the TCP/IP network layer with an IB network layer on top of the Ethernet link layer and does not support IP routing. The Ethernet type is 0x8915. in RoCE, the link layer header of the infiniband is removed and the GUID used to represent the address is converted to an Ethernet MAC. infiniband relies on lossless physical transport, and RoCE relies on lossless Ethernet transport.

RoCEv2

Since the RoCEv1 data frame does not have an IP header, it can only communicate within a 2-tier network. To solve this problem, in 2014 IBTA proposed RoCE V2, which extends RoCEv1 by replacing GRH (Global Routing Header) with a UDP header + IP header. Because RoCE v2 packets are routable at Layer 3, they are sometimes referred to as "Routable RoCE" or "RRoCE" for short. As shown in the figure below.

RoCE technology can be implemented through a common Ethernet switch, but the server needs to support RoCE network cards. Since RoCEv2 is a UDP protocol, although the UDP protocol is relatively high efficiency, but unlike the TCP protocol, there is a retransmission mechanism to ensure reliable

Technology	Data Rates (Gbit/s)	Latency	Key Technology	Advantage	Disadvantage
TCP/IP over Ethernet	10, 25, 40, 50, 56, 100, or 200	500-1000 ns	TCP/IP Socket programming interface	Wide application scope, low price, and good compatibility	Low network usage, poor average performance, and unstable link transmission rate
Infiniband	40, 56, 100, or 200	300-500 ns	InfiniBand network protocol and architecture Verbs programming interface	Good performance	Large-scale networks not supported, and specific NICs and switches required
RoCE/RoCEv2	40, 56, 100, or 200	300-500 ns	InfiniBand network layer or transport layer and Ethernet link layer Verbs programming interface	Compatibility with traditional Ethernet technologies, cost-effectiveness, and good performance	Specific NICs required Still have many challenges to
Omni-Path	100	100 ns	OPA network architecture Verbs programming interface	Good performance	Single manufacturer and specific NICs and switches required

Table X - Compares RDMA Network Technologies

transmission, once there is a packet loss, must rely on the upper layer of the application found and then do retransmission, which will greatly reduce the transmission efficiency of RDMA. So in order to play out the true effect of RoCE, it is necessary to build a lossless network environment for RDMA without losing packets.

RoCE can run in both lossless and compromised network environments, called Resilient RoCE if running in a compromised network environment, and Lossless RoCE if running in a compromised network environment.

RDMA is more and more widely used in market, especially in OTT companies. There have been tens of thousands of servers supporting RDMA, carrying our databases, cloud storage, data analysis systems, HPC and machine learning applications in production. Applications have reported impressive improvements by adopting RDMA [7]. For instance, distributed machine learning training has been accelerated by 100+ times compared with the TCP/IP version, and the I/O speed of SSD-based cloud storage has been boosted by about 50 times compared to the TCP/IP version. These improvements majorly stem from the hardware offloading characteristic of RDMA.

4 Challenges with today's data center network

High bandwidth and low latency tradeoff

✓ It's difficult to achieve high bandwidth and low latency simultaneously

- ✓ Experimentation shows the tradeoff still exists after varying algorithms, parameters, traffic patterns and link loads
- ✓ Reason explanation about why tradeoff exists

Deadlock free lossless network

- ~~✓ High-performance RDMA applications requires lossless network (Zero packet loss and low latency)~~
- ~~✓ Lossless Ethernet requires Priority-based Flow Control (PFC, in IEEE802.1Qbb)~~
- ~~✓ PFC storm may cause severe deadlock problem in data center~~
- ~~✓ Example deadlock problem in a CLOS network~~

As mentioned above, RDMA has advantages such as low latency, high throughput, and low CPU usage compared with TCP. Unlike TCP, RDMA needs a lossless network; i.e. there must be no packet loss due to buffer overflow at the switches [14]. The RoCE protocol is based on UDP and requires PFC to ensure that no packet loss occurs on the entire network. Otherwise, packet loss on the network will severely affect service performance. As shown in figure X, the RoCE service throughput decreases rapidly with the increase of the packet loss rate. Therefore, even if one thousandth of the packet is lost on the network, the service performance decreases by about 30%.

Priority-based Flow Control (IEEE Std 802.1Q-2018, Clause 36 [14]) prevents buffer overflow by pausing the upstream sending entity when buffer occupancy exceeds a specified threshold. However, there are some problems with the large-scale use of PFC. Enabling PFC on the entire network can prevent packet loss. However, a PFC deadlock may occur. Once the deadlock occurs, service traffic is interrupted, causing great harm. In addition, even if the traffic stops, the deadlock still cannot be automatically removed.

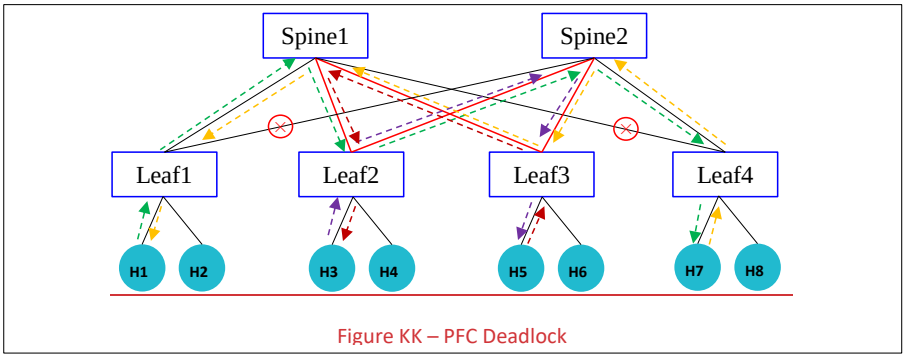


Figure KK – PFC Deadlock

The red cross indicates a transient or permanent fault in Figure X, such as link failure, port failure, or route failure. The server sends traffic from H1 to H7 (green line) and from H3 to H5 (purple line). The failure causes flows from H1 to H7 reroute between leaf 2 and leaf 3. There is a backward '8' shape traffic between spine 1, spine 2, leaf 2 and leaf 3. The Cyclic Buffer Dependency (CBD) is formed now. If the buffer depth of the corresponding port in the CBD reaches the Xoff threshold, a PFC deadlock occurs.

When the network scale is small, the PFC deadlock probability is low. However, with the high-performance service requirements of data centers and the RoCE protocol is widely used, the

network scale will be larger and larger. As the scale increases, the probability of PFC deadlock increases exponentially, which severely affects user experience and is unacceptable for the service SLA.

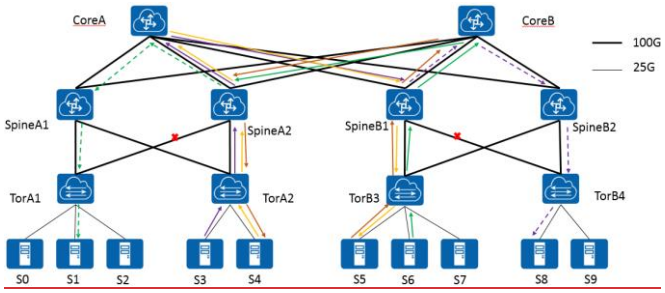
Name	PFC deadlock test
Target	proves that the PFC Deadlock problem occurs on a large-scale network
Test Environment	Test topology：  Prerequisites： 1） All the devices work in the normal state； 2） Set up the test environment according to the preceding networking diagram。
Test Steps	1） Enable PFC in queues 4 and 5； 2） Queue 4 and queue 5 are scheduled in WRR mode. The weights of queue 4 and queue 5 are the same.； 3） Layer 2 interface: trust dscp； 4） When there are four flows in network: S3 to S8 (purple), S6 to S1 (green), S4 to S5 (orange), S5 to S4 (red), the CBD we mention below occurs; 5） Two links are shut down transiently (red cross in the figure);
Expected Results	1） A PFC deadlock occurs. 2） Stop all flows transmission， PFC deadlock still persists.

Table ZZZ – PFC Deadlock Test

Table ZZZ shows the result of test. It shows that PFC deadlock situation occurs. In this situation, the bandwidth and transmission speed of network come to nearly zero.

```
(l): loopback
(s): spoofing
(b): BFD down
(e): ETHOAM down
(d): Dampening Suppressed
(p): port alarm down
(dl): DLDP down
(c): CFM down
InUti/OutUti: input utility rate/output utility rate
Interface                    PHY      Protocol  InUti OutUti   inErrors  outErrors
100GE1/3/1                   up       up        0.05%     0%          0          0
100GE1/3/2                   up       up           0%  0.05%          0          0
Eth-Trunk1                   up       up           0%  0.05%          0          0
  100GE1/1/1:1(25GE)         up       up           0%  0.05%          0          0
  100GE1/1/2:1(25GE)         up       up           0%  0.01%          0          0
Eth-Trunk2                   up       up           0%  0.01%          0          0
  100GE1/1/3:1(25GE)         up       up           0%  0.01%          0          0
  100GE1/1/4:1(25GE)         up       up           0%  0.01%          0          0
Eth-Trunk3                   up       up        0.01%  0.01%          0          0
  100GE1/1/5:1(25GE)         up       up        0.01%     0%          0          0
  100GE1/1/6:1(25GE)         up       up           0%  0.01%          0          0
Eth-Trunk4                   up       up           0%  0.01%          0          0
  100GE1/1/7:1(25GE)         up       up           0%  0.01%          0          0
  100GE1/1/8:1(25GE)         up       up           0%  0.01%          0          0
Eth-Trunk5                   up       up           0%  0.01%          0          0
  100GE1/2/1:1(25GE)         up       up           0%  0.01%          0          0
```

Figure SSS – Evidence of PFC deadlock

```
GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:167:05:39
remote address: LID 0000 QPN 0x11b9 PSN 0x1ff1a9 RKey 0x02dc55 VAddr 0x007f79d1c7
GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:167:05:39
remote address: LID 0000 QPN 0x11ba PSN 0xeade36 RKey 0x031ee4 VAddr 0x007f195e2b
GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:167:05:39
remote address: LID 0000 QPN 0x11b4 PSN 0x80b750 RKey 0x029bf0 VAddr 0x007f05c62e
GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:167:05:39
remote address: LID 0000 QPN 0x11b0 PSN 0xfe38f1 RKey 0x0279d1 VAddr 0x007ff87445
GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:167:05:39
---------------------------------------------------------------------------------------
#bytes     #iterations    BW peak[Gb/sec]    BW average[Gb/sec]   MsgRate[Mpps]
---------------------------------------------------------------------------------------
#bytes     #iterations    BW peak[Gb/sec]    BW average[Gb/sec]   MsgRate[Mpps]
---------------------------------------------------------------------------------------
#bytes     #iterations    BW peak[Gb/sec]    BW average[Gb/sec]   MsgRate[Mpps]
---------------------------------------------------------------------------------------
#bytes     #iterations    BW peak[Gb/sec]    BW average[Gb/sec]   MsgRate[Mpps]
---------------------------------------------------------------------------------------
#bytes     #iterations    BW peak[Gb/sec]    BW average[Gb/sec]   MsgRate[Mpps]
8192       6800             0.00               0.89                 0.013631
8192       5900             0.00               0.78                 0.011827
8192       0            0.000000           0.000000             0.000000
8192       0            0.000000           0.000000             0.000000
8192       0            0.000000           0.000000             0.000000
Completion with error at client
Failed status 12: wr_id 0 syndrom 0x81
Completion with error at client
scnt=5828, ccnt=5828
Failed status 12: wr_id 0 syndrom 0x81
scnt=5728, ccnt=5728
Error occurred while running infinitely! aborting ...
Error occurred while running infinitely! aborting ...
Completion with error at client
```

Figure SSSS – More Evidence of PFC deadlock

Congestion control issues in large-scale data center networks

- ✓ How large scale today's data center is?
- ✓ Use cases for TCP and RoCE flows mixture
- ✓ Smart-buffer mechanisms in mainstream switch chips
- ✓ SLAs cannot be guarantee when TCP and RoCE traffic coexists

Configuration complexity of congestion control algorithms

- ✓ Tuning RDMA networks is an important factor to achieving high-performance
- ✓ Current method of parameters configuration can be a complex operation
- ✓ Congestion control algorithms usually requires collaboration between the NIC and switch
- ✓ Traditional PFC manual configuration needs complex calculation with lots of parameters
- ✓ Excessive headroom leads to reduce the number of lossless queues while too little headroom leads to packet loss

5 New technologies to address new data center problems

Approaches to PFC storm elimination

- ✓ Tuning RDMA networks is an important factor to achieving high-performance
- ✓ Current method of parameters configuration can be a complex operation
- ✓ Congestion control algorithms usually requires collaboration between the NIC and switch
- ✓ Traditional PFC manual configuration needs complex calculation with lots of parameters
- ✓ Excessive headroom leads to reduce the number of lossless queues while too little headroom leads to packet loss

Improving Congestion Notification

- ✓ Improved Explicit Congestion Notification
- ✓ Enhanced version of Quantized Congestion Notification (originally IEEE 802.1Qau)
- ✓ Intelligent Methods of improving QoS support in mixed traffic environments
- ✓ Test verification (ODCC lossless DCN test specification and result)

Intelligent congestion parameter optimization

- ✓ Intelligent heuristic algorithms for identifying congestion parameters
- ✓ Methods for dynamic optimization based on services
- ✓ Test verification (ODCC lossless DCN test specification and result)

Buffer optimization of lossless queues

- ✓ Intelligent headroom calculation

✓ Self-adaptive headroom configuration

6 Standardization Considerations

Things for the IEEE 802 and IETF to consider. Possibly others as well – SNIA, IBTA, NVMe, etc..

7 Conclusion

Closing words…

8 Citations

<< format the table later – MS word screws up the format each time your rebuild, so just wait until the end and change the column widths to get it to look correct. >>

[1] IEEE, "Nendica Work Item: Data Center Networks," [Online]. Available: https://1.ieee802.org/nendica-DCN/. [Accessed 14-05-2020].

[2] IEEE, "IEEE 802 Nendica Report: The Lossless Network for Data Centers," 17-8-2018. [Online]. Available: https://xploreqa.ieee.org/servlet/opac?punumber=8462817. [Accessed 13-05-2020].

[3] J. Handy and T. Coughlin, "Survey: Users Share Their Storage," 12-2014. [Online]. Available: https://www.snia.org/sites/default/files/SNIA%20IOPS%20Survey%20White%20Paper.pdf. [Accessed 14-05-2020].

[4] Huawei, "AI, This Is the Intelligent and Lossless Data Center Network You Want!," 13-March 2019. [Online]. Available: https://www.cio.com/article/3347337/ai-this-is-the-intelligent-and-lossless-data-center-network-you-want.html. [Accessed 14-05-2020].

[5] E. K. Karuppiah, "Real-World Problem Simplification Using Deep Learning / AI," 2-November 2017. [Online]. Available: https://www.fujitsu.com/sg/Images/8.3.2%20FAC2017Track3_EttikanKaruppiah_RealWorldProblemSimplificationUsingDeepLearningAI%20.pdf. [Accessed 14-05-2020].

[6] O. Cardona, "Towards Hyperscale High Performance Computing with RDMA," 12 June 2019. [Online]. Available: https://pc.nanog.org/static/published/meetings/NANOG76/1999/20190612_Cardona_Towards_Hyperscale_High_v1.pdf. [Accessed 14 05 2020].

[7] Y. Li, R. Miao, H. H. Liu, Y. Zhuang, F. Feng, L. Tang, Z. Cao, M. Zhang, F. Kelly, M. Alizadeh and M. Yu, "HPCC: high precision congestion control," in *Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM '19)*, New York, NY, USA, 2019.

[8] J. L. Jacobi, "NVMe SSDs: Everything you need to know about this insanely fast storage," 10 March 2019. [Online]. Available: https://www.pcworld.com/article/2899351/everything-you-need-to-know-about-nvme.html. [Accessed 14 05 2020].

[9] M. Alipio, N. M. Tiglao, F. Bokhari and S. Khalid, "TCP incast solutions in data center networks: A classification and survey," *Journal of Network and Computer Applications,* vol. 146, p. 102421, 2019.

[10] T. P. Morgan, "Machine Learning Gets An Infiniband Boost With Caffe2," 19 April 2017. [Online]. Available: https://www.nextplatform.com/2017/04/19/machine-learning-gets-infiniband-boost-caffe2/. [Accessed 14 05 2020].

[1] IEEE, "Nendica Work Item: Data Center Networks," [Online]. Available: https://1.ieee802.org/nendica-DCN/. [Accessed 14 05 2020].

[2] IEEE, "IEEE 802 Nendica Report: The Lossless Network for Data Centers," 17 8 2018. [Online]. Available: https://xploreqa.ieee.org/servlet/opac?punumber=8462817. [Accessed 13 05 2020].

[3] J. Wiles, "Mobilize Every Function in the Organization for Digitalization," Gartner, 03 December 2018. [Online]. Available: https://www.gartner.com/smarterwithgartner/mobilize-every-function-in-the-organization-for-digitalization/. [Accessed 10 June 2020].

[4] Huawei, "Huawei Predicts 10 Megatrends for 2025," Huawei, 08 August 2019. [Online]. Available: https://www.huawei.com/en/press-events/news/2019/8/huawei-predicts-10-megatrends-2025. [Accessed 10 June 2020].

[5] J. Handy and T. Coughlin, "Survey: Users Share Their Storage," 12 2014. [Online]. Available: https://www.snia.org/sites/default/files/SNIA%20IOPS%20Survey%20White%20Paper.pdf. [Accessed 14 05 2020].

[6] Huawei, "AI, This Is the Intelligent and Lossless Data Center Network You Want!," 13 March 2019. [Online]. Available: https://www.cio.com/article/3347337/ai-this-is-the-intelligent-and-lossless-data-center-network-you-want.html. [Accessed 14 05 2020].

[7] E. K. Karuppiah, "Real World Problem Simplification Using Deep Learning / AI," 2 November 2017. [Online]. Available: https://www.fujitsu.com/sg/Images/8.3.2%20FAC2017Track3_EttikanKaruppiah_RealWorldProblemSimplificationUsingDeepLearningAI%20.pdf. [Accessed 14 05 2020].

[8] O. Cardona, "Towards Hyperscale High Performance Computing with RDMA," 12 June 2019. [Online]. Available: https://pc.nanog.org/static/published/meetings/NANOG76/1999/20190612_Cardona_Towards_Hyperscale_High_v1.pdf. [Accessed 14 05 2020].

[9] J. L. Jacobi, "NVMe SSDs: Everything you need to know about this insanely fast storage," 10 March 2019. [Online]. Available: https://www.pcworld.com/article/2899351/everything-you-need-to-know-about-nvme.html. [Accessed 14 05 2020].

[10] M. Alipio, N. M. Tiglao, F. Bokhari and S. Khalid, "TCP incast solutions in data center networks: A classification and survey," *Journal of Network and Computer Applications*, vol. 146, p. 102421, 2019.

[11] Z. Jai, Y. Kwon, G. Shipman, P. McCormick, M. Erez and A. Aiken, "A distributed multi-GPU system for fast graph processing," in *VLDB Endownment*, 2017.

[12] T. P. Morgan, "Machine Learning Gets An Infiniband Boost With Caffe2," 19 April 2017. [Online]. Available: https://www.nextplatform.com/2017/04/19/machine-learning-gets-infiniband-boost-caffe2/. [Accessed 14 05 2020].

[13] Y. Li, R. Miao, H. H. Liu, Y. Zhuang, F. Feng, L. Tang, Z. Cao, M. Zhang, F. Kelly, M. Alizadeh and M. Yu, "HPCC: high precision congestion control," in *Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM '19)*, New York, NY, USA, 2019.

[14] C. Guo, H. Wu, Z. Deng, G. Soni, J. Ye, J. Padhye and M. Lipshteyn, "RDMA over Commodity Ethernet at Scale," in *In Proceedings of the 2016 ACM SIGCOMM Conference (SIGCOMM '16)*, 2016.

[15] IEEE, IEEE Std 802.1Q-2018, IEEE Standard for Local and Metropolitan Area Networks — Bridges and Bridged Networks, IEEE, 2018.