ODCC
开放数据中心委员会

冬季全会
Winter Plenary

PHYSec安全加密技术及在智算中心的应用

演讲人：王豪杰
演讲单位：中国移动研究院

2023

目录

CONTENTS

01 场景与安全需求

02 现有技术分析

03 PHYSec技术思路

04 总结与展望

01

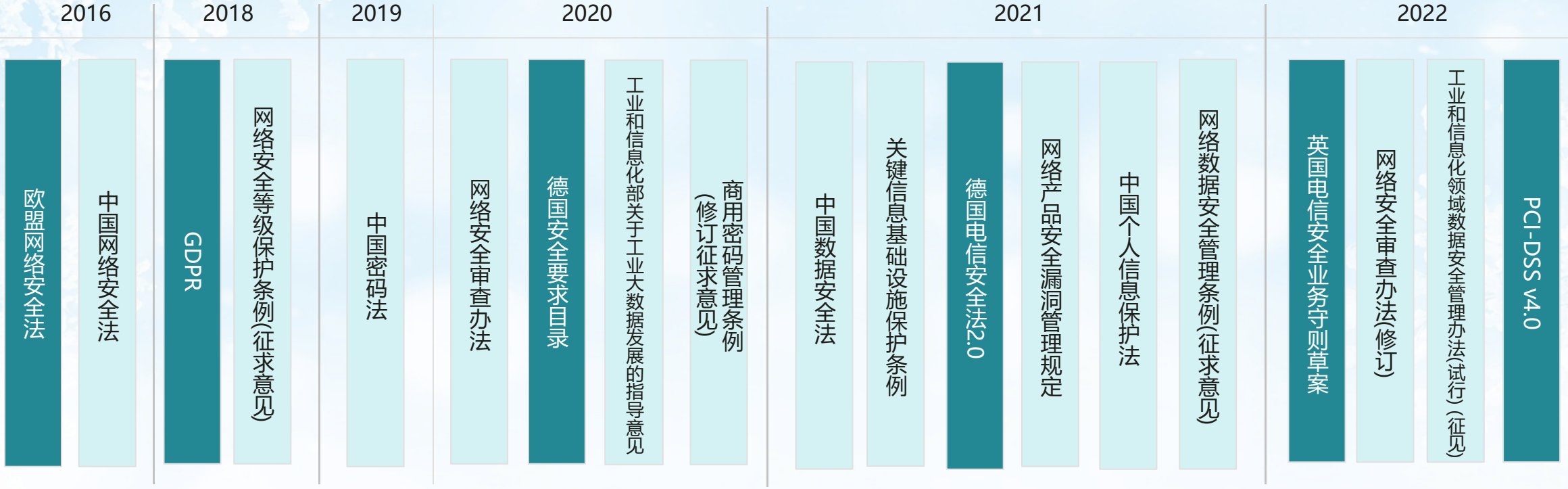
Chapter one

场景与安全需求

网络链路攻击事件不断被爆出，全球加快数据安全治理

ODCC 开放数据中心委员会

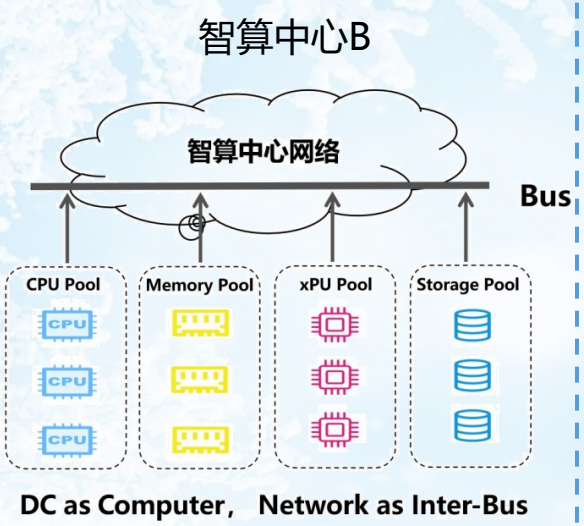
网络攻击日渐频繁，无处不在! 在Internet上每天有超过1亿次的网络攻击，每秒超1000次以上的攻击发生。攻击者企图访问用户个人或组织的信息数据，没有一家公司能幸免

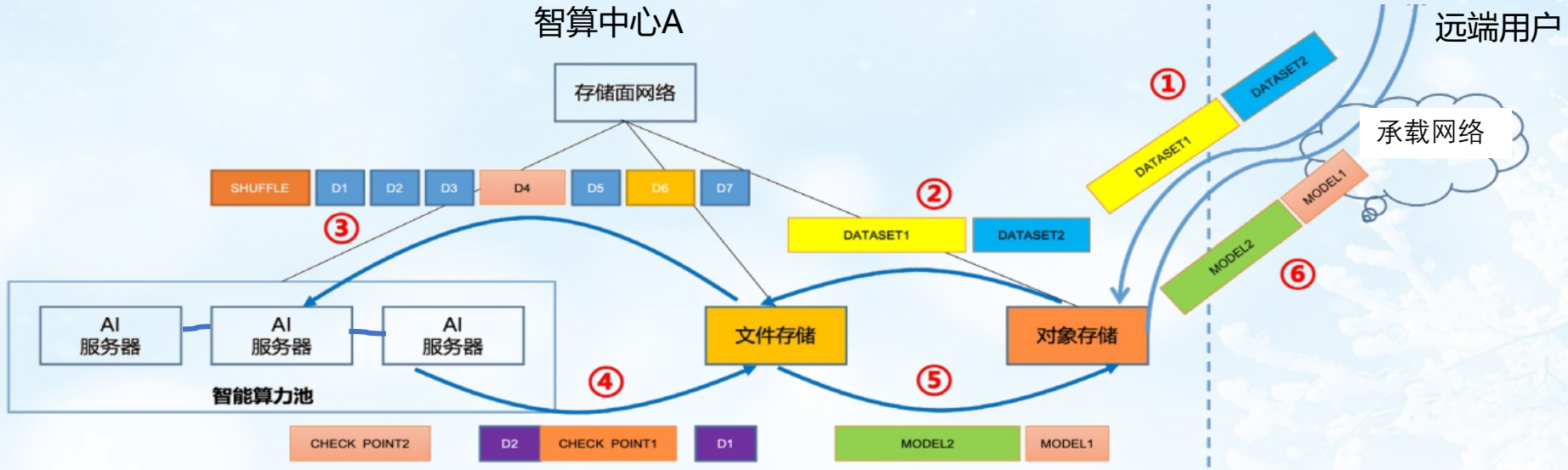


全球100多个国家已对数据安全提出要求；中国、英国、德国、欧盟相继颁布网络安全法案，要求通信网络确保数据传输安全

智算中心的数据安全需求

用户入算上载样本数据、算内及算间训练生成模型及参数、下载训练后的模型及参数，涉及敏感数据的频繁交互，存在泄露、窃听的风险





算间网络

⑦**参数传递**：端到端安全加密，确保模型框架、参数安全

算内网络

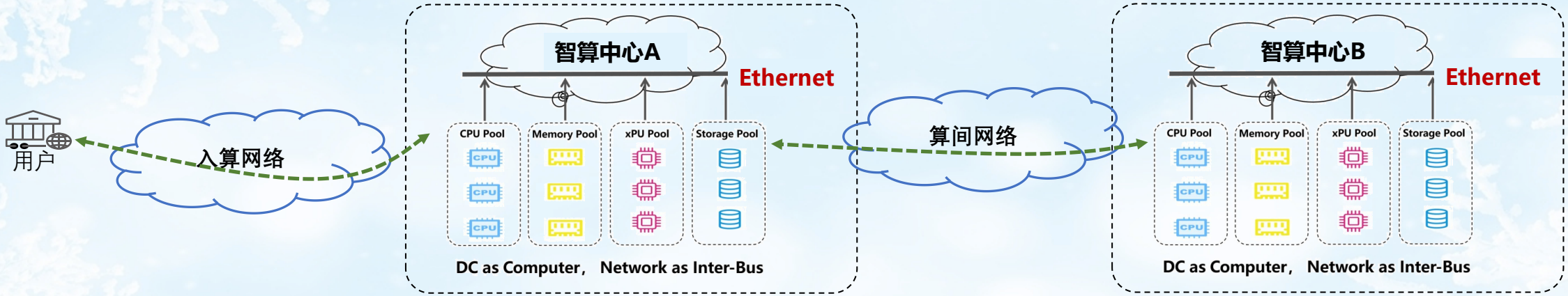
②**数据转移**：存储节点直接(对象➔文件)，需确保样本数据安全
③**数据读取**：计算节点访问文件存储节点(读)，需确保样本数据安全
④**归档写回**：计算节点访问文件存储节点(写)，需确保模型与参数安全
⑤**模型复制**：存储节点直接(文件➔对象)，需确保模型与参数安全

入算网络

①**数据上传**：端到端安全加密，确保样本数据安全
⑥**模型下载**：端到端安全加密，确保模型框架、参数安全

算内“低时延高吞吐”，入算、算间“高安全低开销”

智算中心内实现网络安全机制的从无到有，入算网络和算间网络实现安全防护能力的由低到高；面向算内、入算和算间的不同诉求，研究合适的安全加密机制



入算关键诉求：
- 用户访问以及用户数据的安全

关键能力：
- 高安全、低开销

算内关键诉求：
- AI模型架构、参数核心资产的存储、使用安全

关键能力：
- 低时延、高吞吐、低开销安全，不占用算力，不影响算效

算间关键诉求：
- 数据与协议安全，防窃听

关键能力：
- 高安全、高吞吐、低开销

上述智算中心网络场景的底层承载网络主流技术是以太网，为了应对上述日益严峻的数据安全挑战，须对以太网提供安全认证、密钥管理以及数据加解密能力，构筑以太网安全机制，为新型智算中心提供安全数据传输能力

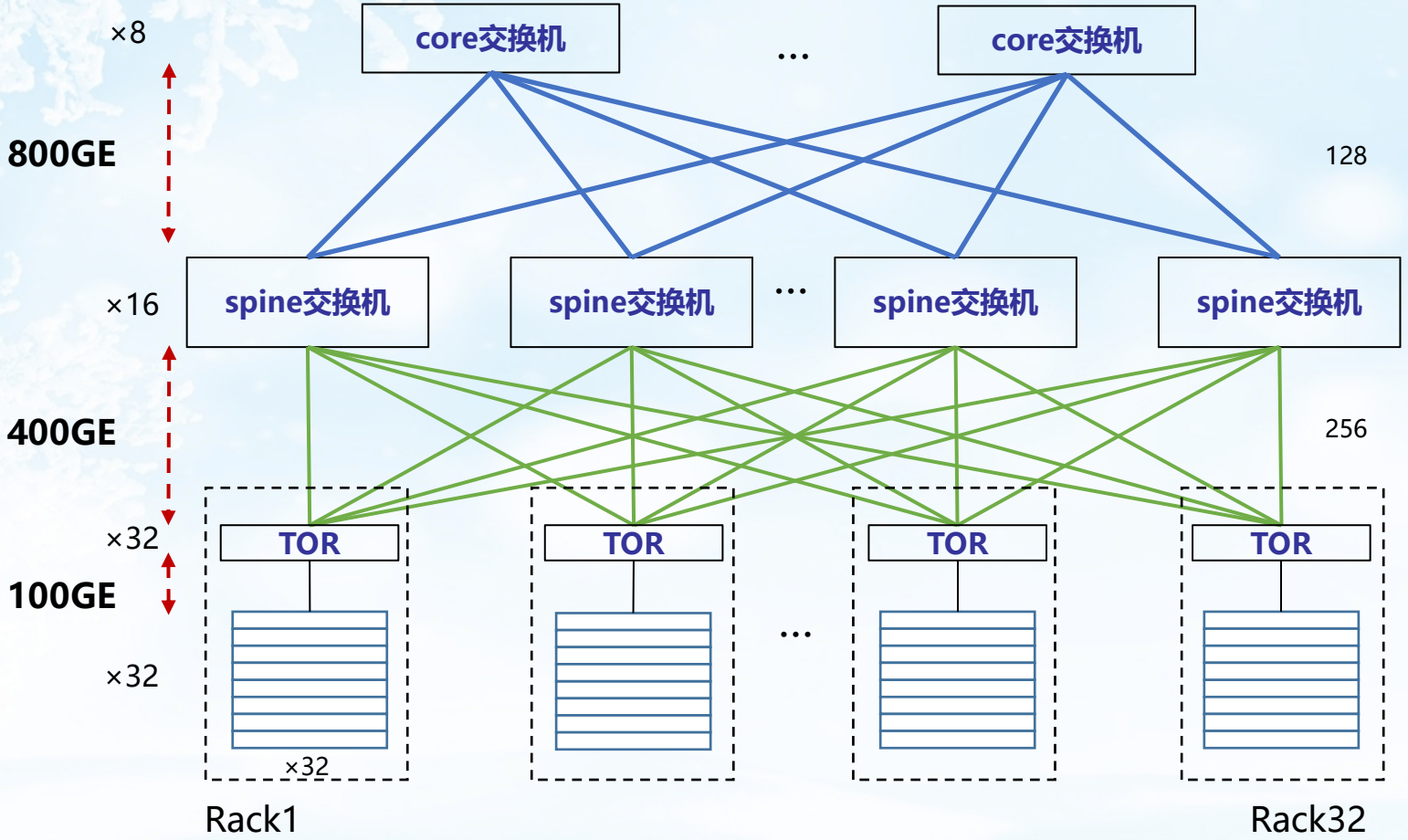
02

Chapter one

现有技术分析

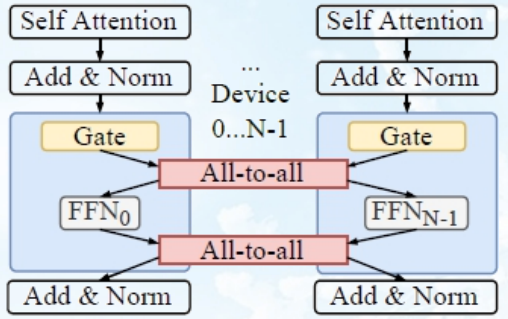
智算中心典型组网

智算中心采用CLOS组网来满足日益增长的转发规模需求，通用性有明显优势。智算中心的AI典型Leaf-Spine组网，ALL-TO-ALL通讯，100GE/400GE/800GE接口，1K计算节点(32×32)

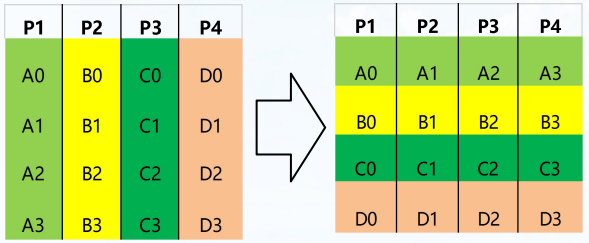


训练和推理：**

- 模型：Mixture of Experts (MoE)



- 通讯方案：all-to-all



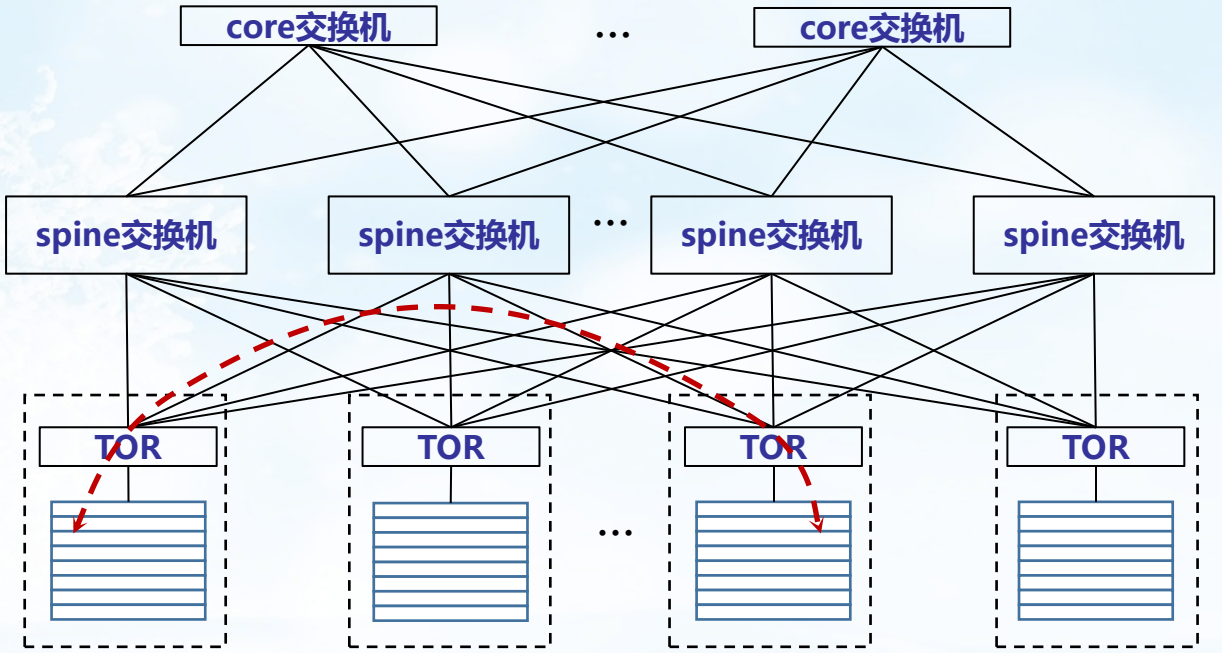
*** 2023 usenix， Accelerating Distributed MoE Training and Inference with Lina，HKU*

TLS/RDMASec/IPSec技术分析

TLS/RDMASec*/IPSec等端到端安全加密方案，服务器间实施典型的ALL-TO-ALL通讯方案；但没有网络的参与，计算节点之间部署端到端方案成本、难度和算力代价较大

优势

- 可以实现计算节点与计算节点，计算节点与存储节点，存储节点与存储节点之间的安全加密



劣势

- 带宽开销增加至少21字节：
  - TLS逐包增加21B开销，RDMASec逐包增加40B开销，IPSec逐包增加48B开销
- 须升级硬件支持：
  - 在AI计算场景下，将TLS/IPSec/PSP安全加密功能卸载到硬件，需要Server硬件支持
- 管控复杂 $O(N^2)$：N台Server
  - ✓ 全网需维护C_N^2安全连接（SA会话）
  - ✓ 每连接1对密钥，全网管理维护$2 \times C_N^2$个密钥
- 静态时延增加超30%：
  - ✓ 以4 Hops为例，server-to-server典型静态链路时延 $\approx 2.85\mu s$
  - ✓ 考虑TLS/RDMASec/IPSec加解密时延最优性能1μs，静态时延增加= $1us/2.85\mu s \approx 35\%$

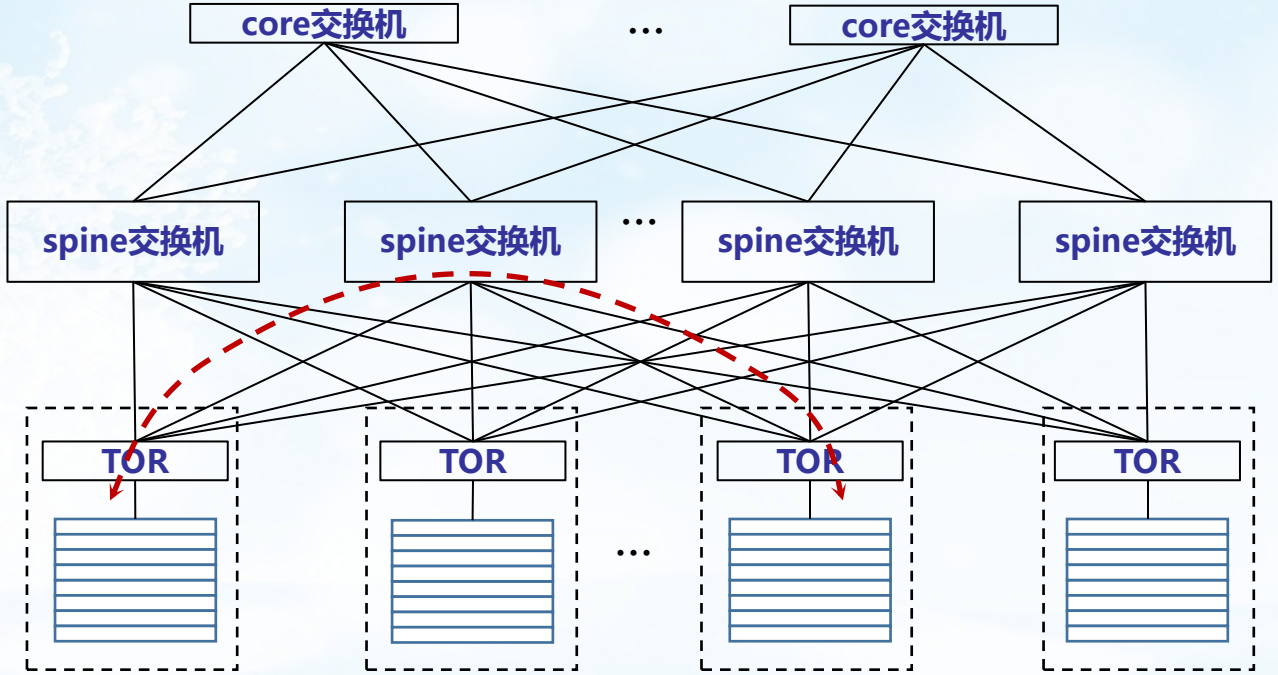
** RFC5042提出RDMASec基于IPSec改进，Google在2022年发布基于IPSec ESP的改进方案PSP*

MACSec技术分析

智算中心内采用MACSec链路级安全加密方案*，服务器间实施典型的ALL-TO-ALL通讯方案；MACSec在管控复杂度有改进，但10多年前主要面向传统园区设计，时延开销无法满足智算时代需求

优势

- **管控复杂度低 $O(N)$:**
  - ✓ Server-TOR, TOR-Spine, Spine-Core部署MACSec，全网维护= **44N**安全连接（SA会话）



劣势

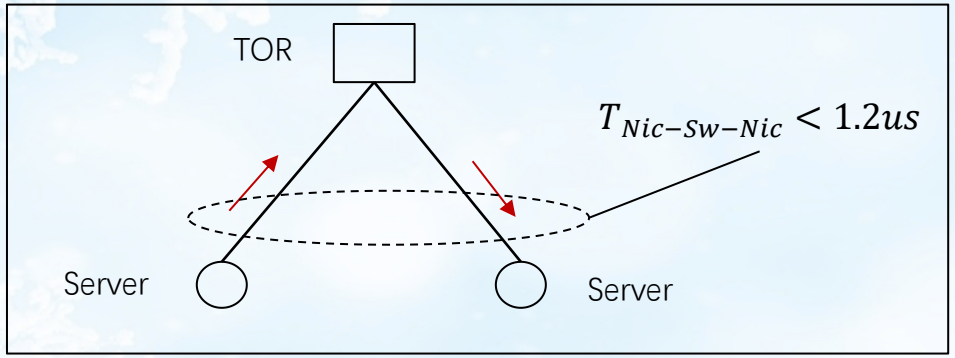
- 带宽开销增加至少32字节:
  - 对64B包而言，逐包开销增加超30%
- 须升级硬件支持:
  - 在AI计算场景下，须将MACSec安全加密功能卸载到硬件，需要Server升级硬件支持，需要交换机升级硬件支持
- 静态时延增加32%：4 Hops
  - ✓ 以4 Hops为例，Server-to-server典型静态时延 $\approx \mathbf{2.85}\mu s$
  - ✓ Server-to-server使能MACSec时典型静态时延 $\approx 3.77\mu s$
  - ✓ 静态时延增加= $(3.77 - 2.85)/2.85 \approx 32\%$

*MACSec2017版本， IEEE802.1AEcg支持VLAN in clear端到端加密方案，此方案优劣势同 TLS/RDMASec/IPSec

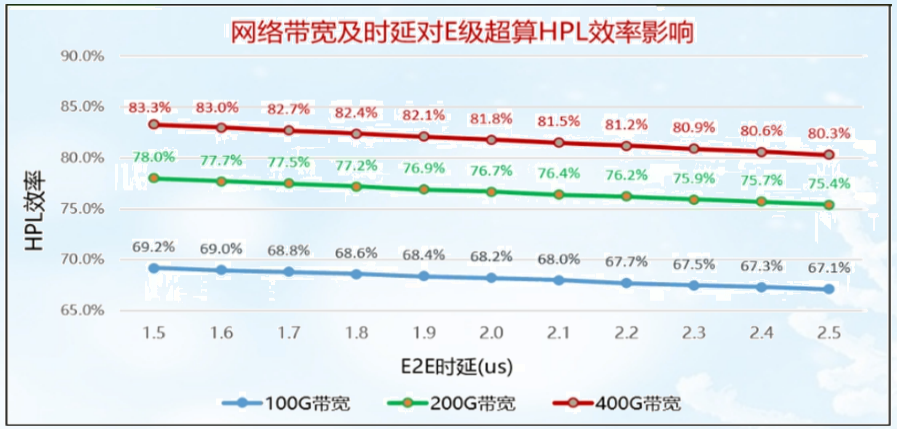
新安全机制：低时延、低开销、配置简单、易部署

设定新安全性能目标，兼顾成本、代价方面，探索智算时代的以太网安全新机制

- **竞争分析：**智算中心最关注算效，通讯时延*对算效影响大
  - ✓ IB控标静态时延 $T_{Nic-Sw-Nic} < 1.2us$



  - ✓ Server-to-Server端到端时延每降低1us，有效算力线性度提升3%



- **各项性能指标对比分析**

最佳　最差　一般

比较项	RDMASec	MACsec	？Sec
加密层级	应用层/传输层	链路层	新层次？
加密开销	带宽开销增加36%	带宽开销增加30%	接近0？
加密时延	静态时延增加35%	静态时延增加32%	静态时延增加< 10%？
加密配置	相对复杂$O(N^2)$	简单$O(N)$	简单$O(N)$？
加载难度	Server升级硬件支持	Server升级硬件支持，交换机升级硬件支持	易升级易部署?

** AI训练和推理，对每一次迭代通讯的尾时延敏感；HPC小包通讯，对RTT时延敏感*

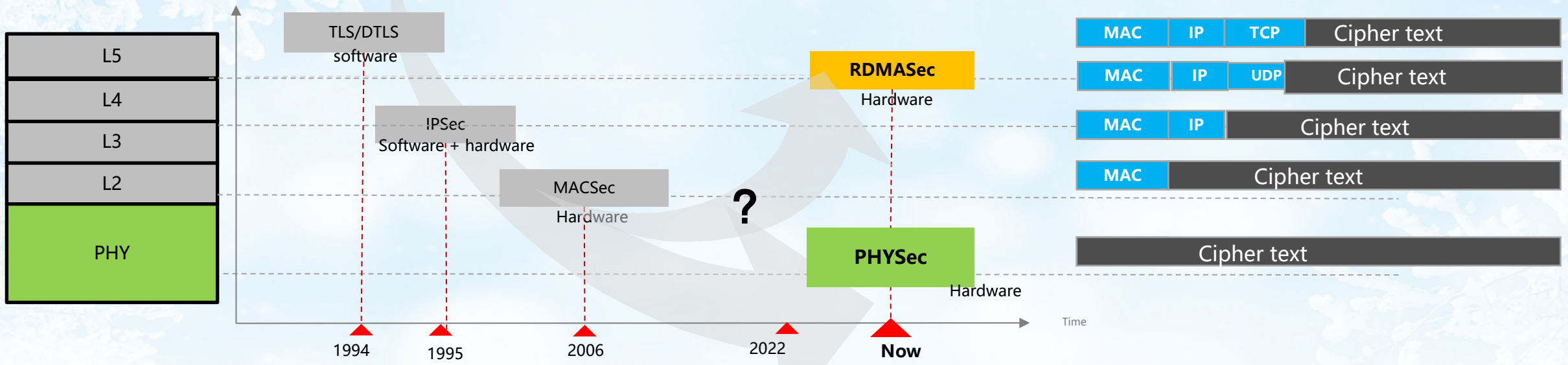
03

Chapter one

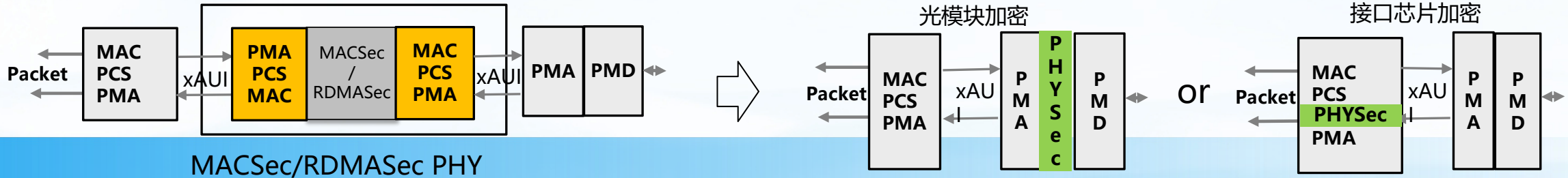
PHYSec技术思路

物理层加密，更低时延、更低开销、协议透明——PHYSec

- 探索新层次：网络安全的思想是把密码学技术应用到网络的不同层次；在智算时代，是完善传统安全机制还是探索“更底层的加解密技术”，解决现有技术方案的安全漏洞与性能瓶颈?

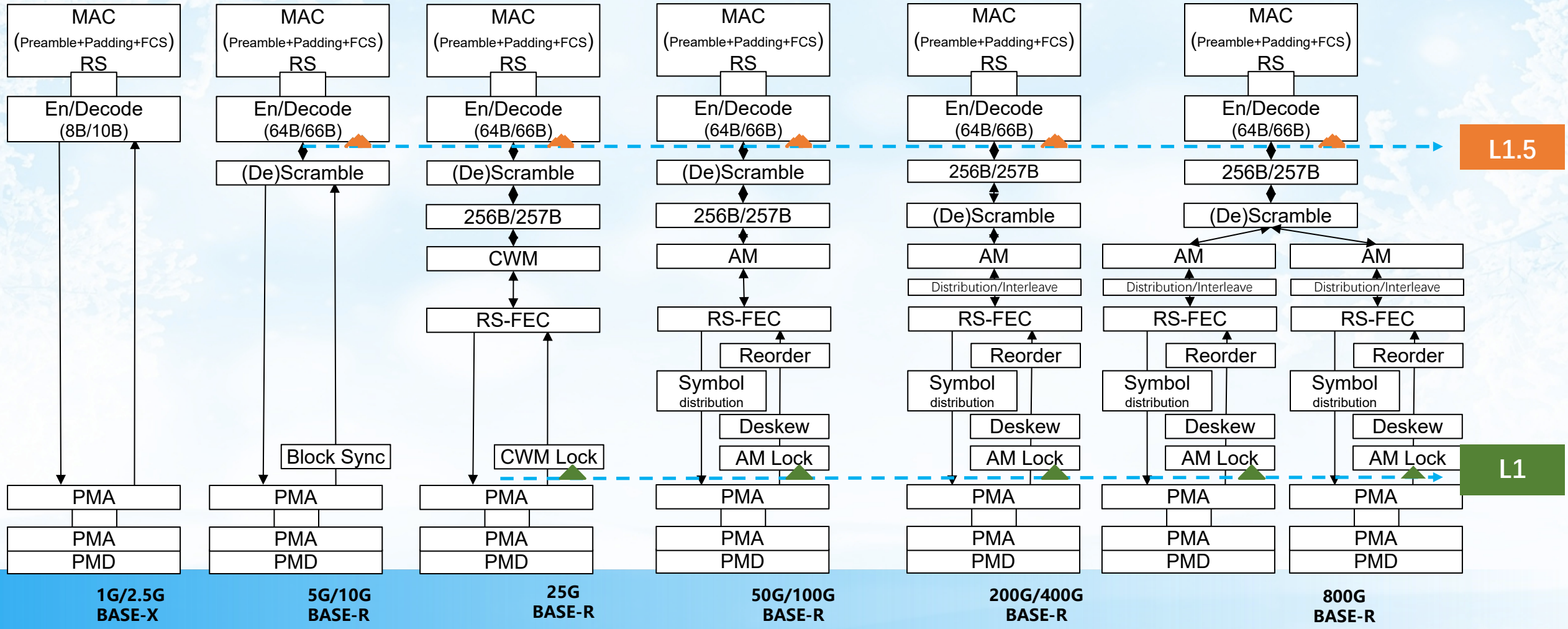


- RDMASec在IPSec和TLS基础上优化，基于硬件实现，优势是能够降低部分时延，但功耗、成本较高，时延仍然影响算效；PHYSec在以太网物理层实现加解密，避免两次背靠背转换，具有极低时延、更低功耗和成本等优势



以太网物理层协议栈实现PHYSec的参考点

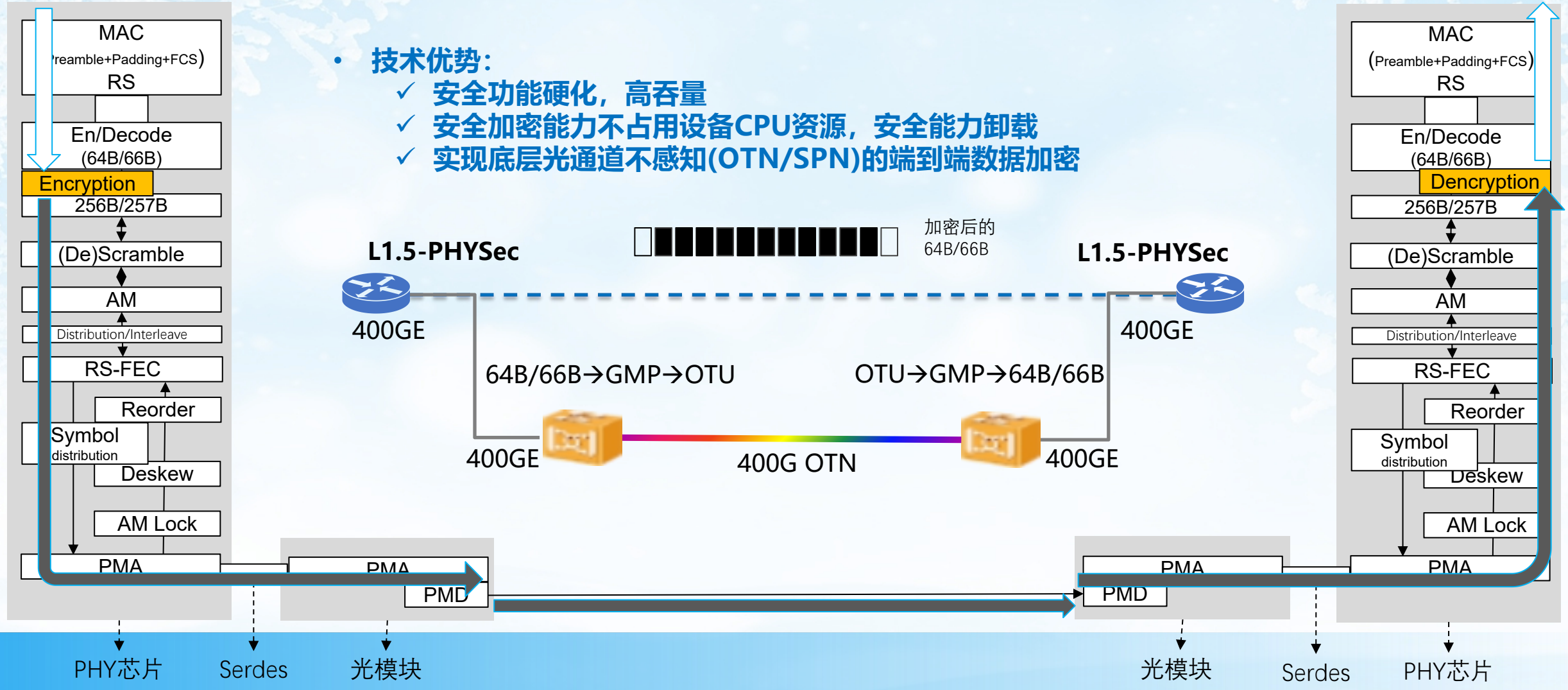
以太网物理层可实现PHYSec参考点：64B/66B码块和virtual lane是两种比较可行的实现位置；不同层次位置实现加解密具有不同的数据防护特点



PHYSec：低开销、低时延、协议透明的高安全线速加解密

基于“码块”加密的L1.5层PHYsec在PHY芯片内实现Encryption/Decryption功能

- **技术优势：**
  - ✓ **安全功能硬化，高吞量**
  - ✓ **安全加密能力不占用设备CPU资源，安全能力卸载**
  - ✓ **实现底层光通道不感知(OTN/SPN)的端到端数据加密**



PHYSec：低开销、低时延、协议透明的高安全线速加解密

基于“比特流”加密的L1层PHYsec在光模块内实现Encryption/Decryption功能

- **技术优势：**
  - ✓ 安全功能可插拔、硬化，高吞量
  - ✓ 无需升级设备硬件，即可具备安全加密能力
  - ✓ 安全加密能力不占用主设备资源，安全能力卸载
  - ✓ 实现端口-端口的链路级数据加解密



PHYSec技术方案的设计原则探讨

☑ **兼容性：PHYSec技术方案在以太网物理层实现，须兼容IEEE802.3以太物理层标准，不影响标准规范的功能与协议**

- 关闭PHYSec，所有功能即原以太网物理层规范的功能；
- 开启PHYSec，PCS或PMA层无感知，PHYSec不修改与上下子层的接口；
- 不修改或占用已有功能协议及协议承载方案，避免后向兼容问题（原有功能未来演进可能会改变既有的使用方式、频率）

☑ **互通性：同一层次方案，要求技术与协议一致，满足互联互通要求。协议承载方案，要求少占用或不占用业务带宽**

- PHYSec原则上可支持通道、链路不同层次的技术方案，类似MACSec（802.1AEbw-2013 for port，802.1AEcg-2017 for channel），实施部署载体可以是PHY接口，也可是光模块或其他载体

☑ **一致性：PHYSec原则上要求采用一套解决方案和协议。协议的承载方式可以根据PHY逻辑子层的要求变化，但要求遵循前述兼容性、互通性**

- 对于IEEE802.3标准规范的100G/200G/400G/800G/1.6T以太接口，物理层及PHY各逻辑子层技术方案虽有差别，但都基于64/66B码块、virtual lane实现，提供了实现方案的一致性基础

ODCC
开放数据中心委员会

04

Chapter one

总结与展望

智算中心安全体系全景

灵活专线

入算

高价值用户

L1.5-PHYSec

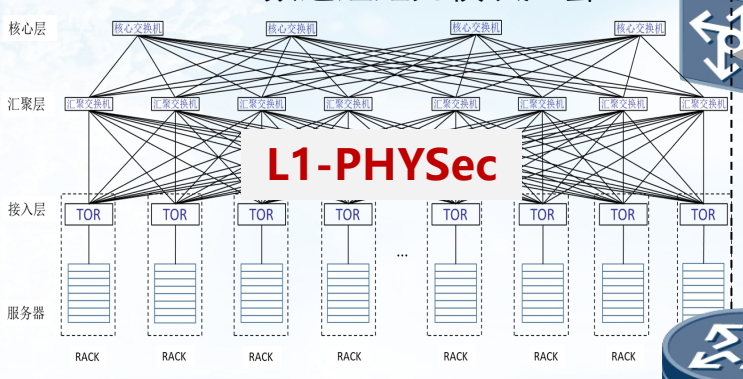
TOP诉求：高安全、低开销

- SPN
- IP/Eth over OTN/WDM/CWMD

智算中心A

算内

数通短距光模块加密



L1-PHYSec

TOP诉求：低时延、高吞吐

算间（短距互联）

L1-PHYSec

数通短距光模块加密

< 40 KM

TOP诉求：高安全、高吞吐

L1.5-PHYSec

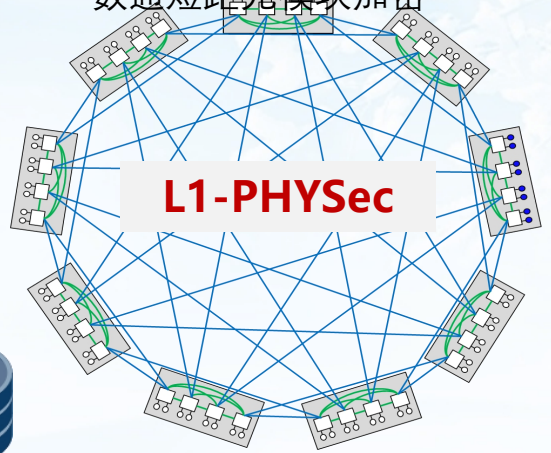
≥ 40KM　　OTN/WDM

算间（长距互联）

智算中心B

算内

数通短距光模块加密



L1-PHYSec

TOP诉求：低时延、高吞吐

总结与展望

目标：在以太网物理层实现低开销、低时延、易配置、易部署的数据安全方案

最佳　最差　一般

比较项	RDMASec	MACsec	PHYSec
加密层级	应用层/传输层	链路层	物理层
加密开销	带宽开销增加36%	带宽开销增加30%	接近0
加密时延	静态时延增加35%	静态时延增加32%	静态时延增加几十ns
加密配置	相对复杂$O(N^2)$	简单$O(N)$	简单$O(N)$
加载难度	Server升级硬件支持	Server升级硬件支持，交换机升级硬件支持	更换光模块易升级易部署

TBD：

1、在新的层次定义加密解密对象，定义帧结构和协议，与上下游生态企业达成共识

2、新机制如何实现接近0的低开销，尽量不占用用户带宽

3、新机制如何实现ns级的时延，基本不影响智算中心的算效?

4、新机制如何保证加解密协议参数的可靠传递，不发生错误

携手业界合作伙伴构建以太网物理层安全体系，助力智算中心安全设施建设的快速发展

ODCC
开放数据中心委员会

谢谢观看

中国移动研究院公众号二维码

中移智库公众号二维码

2023