SYNOPSYS®
Silicon to Software™

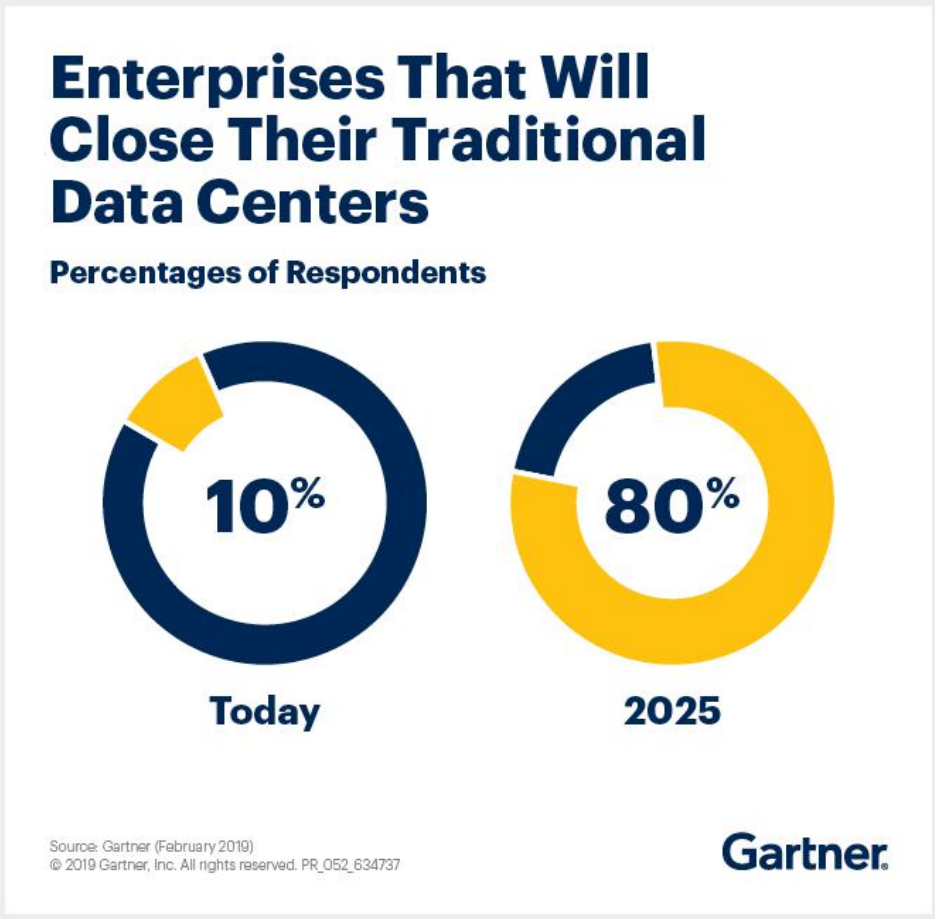
实现PCI Express 5.0和CXL设计的最大吞吐量和最低延迟的关键

张小林，资深产品市场经理

2020年11月10日

外包和“随处访问”推动云计算发展

85%的企业将把他们的数据中心转移到云上，推动新的设计

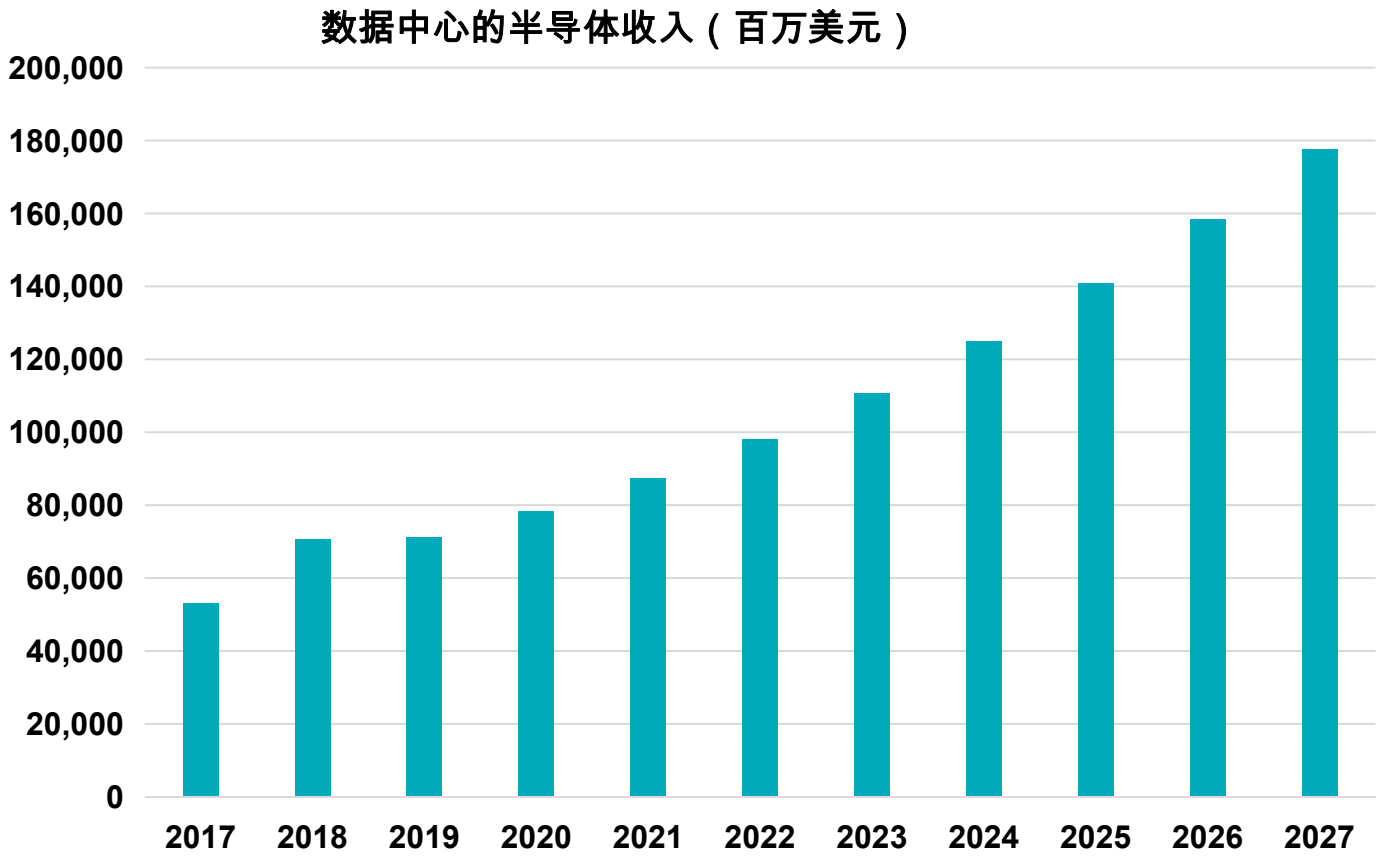


- ❑ 商业、娱乐、游戏、学校和非营利组织大量转向在线服务
- ❑ 新冠疫情时代推动云用量增长，“基本服务”需求旺盛
  - 从3月到4月，Azure不得不增加110 TB的容量和12个新的边缘服务站点
  - Microsoft团队会议记录时长在2周内从9亿分钟/每天增加至27亿分钟/每天
- ❑ 使用Amazon Web services（AWS）、Microsoft Azure、Google Cloud的服务将本地数据中心迁移到云端→推动新设计开始

1：数据中心前沿“全天候提供更多服务器：Microsoft如何应对COVID-19： 2020年6月17日

数据中心推动新的设计起点和创新

数据中心半导体市场将在2027年达到1770亿美元

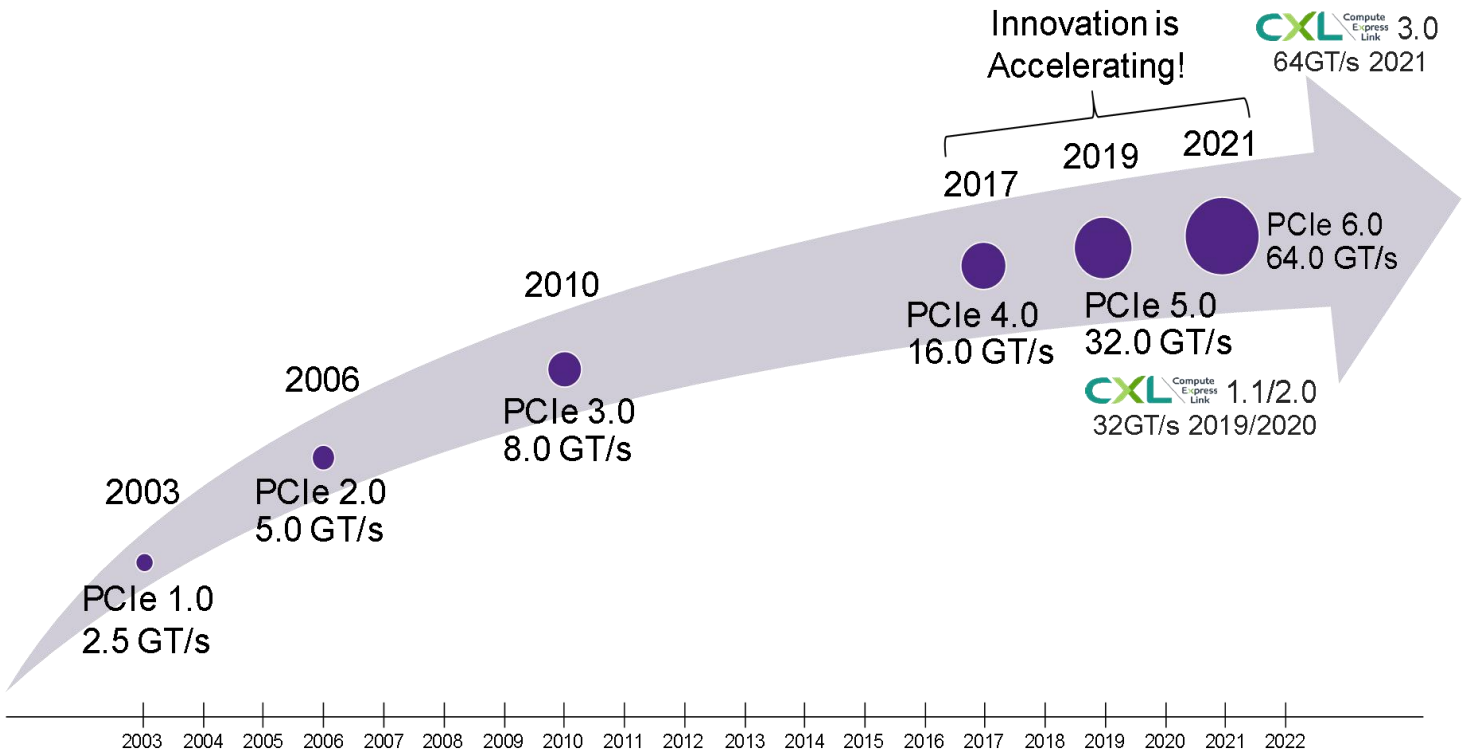


Source: IBS Global Semiconductor Industry Report, 2019

- 2018年数据中心市场为705亿美元(复合年增长率为10.81%)，到2027年预计达1776亿美元。
- 2030年的数据流量预计为2^{60}字节
- 很多公司正在设计基于7nm到3nm的先进工艺芯片，需要大量IP才能通过认证
- OpenAI表示:“在最大规模的人工智能训练中芯片，计算能力呈指数级增长，倍增时间达3.5个月。”

大数据集推动PCIe 5.0的采用

PCIe 4.0标准的延迟加速了对PCIe 5.0的需求

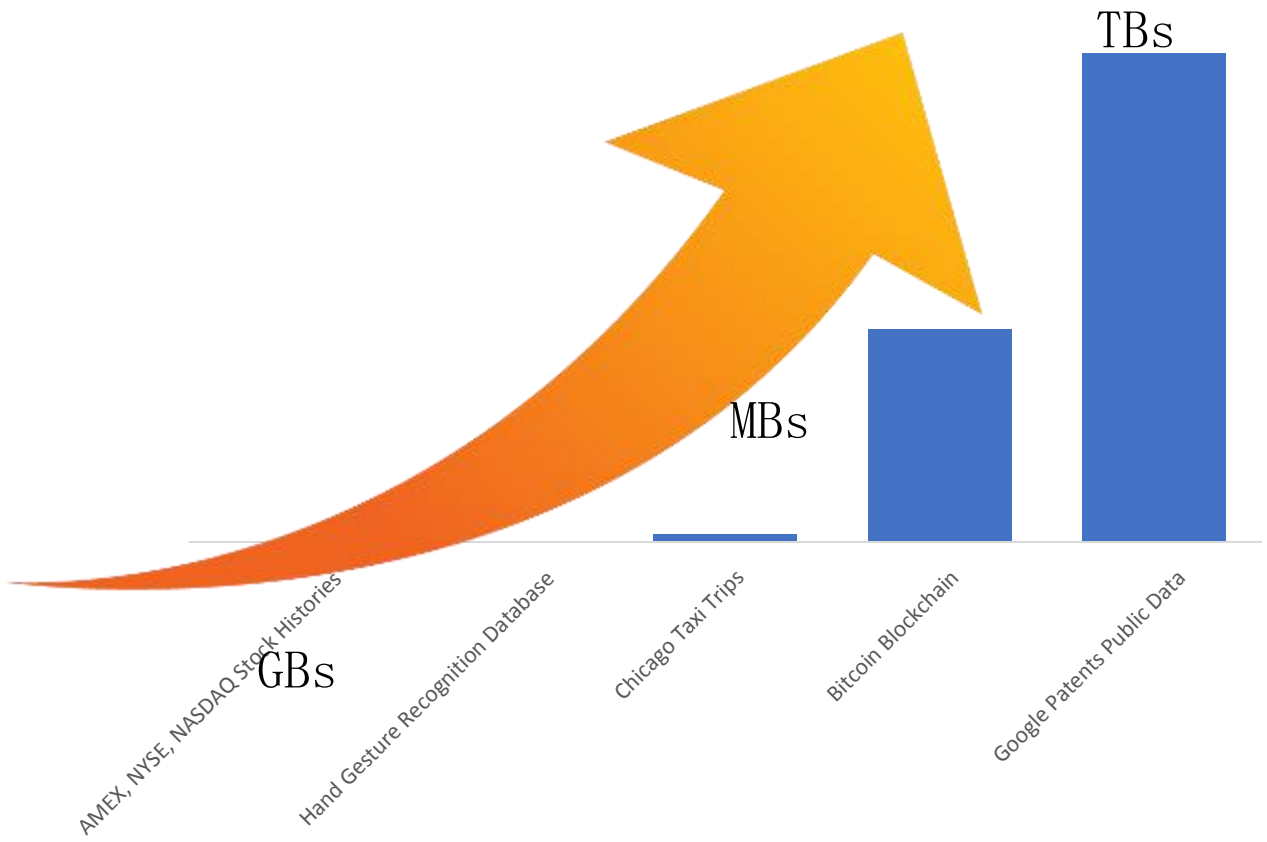


- “稳步向400GbE云网络迈进，以及先进的AI/ML工作负载的发展，推动了PCIe带宽每两年翻一番的需求，以实现数据在计算节点之间的高效移动。”
- PCIe 5.0被迅速推出，来解决400GbE应用带宽的问题
- 巨额投资支持PCIe 5.0生产
- PCIe 5.0 预计在2021年大量部署于CPU应用

[1]*“Accelerating AI And ML Applications With PCIe 5”, Semiconductor Engineering, Jan 2020*

AI/ML和其他数据应用的大小驱动了新的接口需求

CXL的出现是由于大数据的应用导致了更大、更高效、缓存一致性存储的要求



- CXL提供缓存一致性和内存池，在分析期间停止复制大型数据集
- 带CXL功能的 SCM/ Persistent Memory
  - 提供高容量、高速率的数据存储
  - 消除数据的保存/恢复处理时间
  - 为其他用途释放DDR插槽
- 带CXL 2.0 功能的交换机将利用AI加速器实现广泛应用。

Synopsys高性能计算的完整解决方案

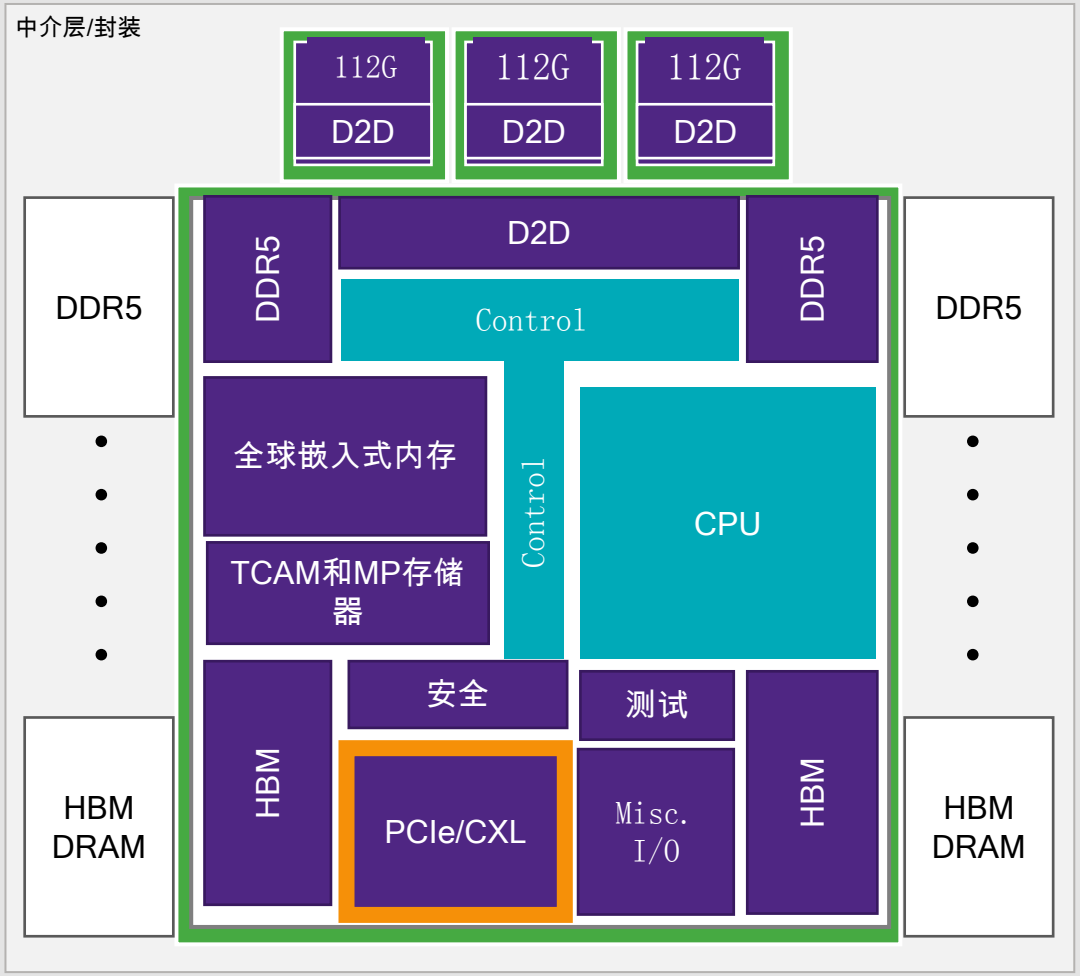
性能、延迟、内存、连接性推动创新

用于背板连接的高速SerDes
400/800G →1.6TB
延迟和功耗是重点

高性能DDR5
具有高带宽和低延迟特性

用于服务器连接的PCIe 5.0。
迄今为止，由于采用的飙升，已发放了100多个许可证

CXL势头强劲
CXL 2.0规范现为版本0.9
CXL 3.0与 PCIe 6.0保持同步
CCIX与CXL融合



并行和串行裸片到裸片方案将满足特定应用的需求而共存。 许多节点需要USR（56G）、XSR（112G）和HBI

最大密度SRAMs，减少芯片面积;
高速度的TCAM支持快速低功耗搜索

硬件信任（ Hardware root of trust ）
TRNG和独特的测试算法多裸片方案

图形、网络和AI驱动下的HBM
HBM2e 3600 →HBM3 6400

适用于高性能计算设计的PCIe 5.0与CXL 2.0

提供在32 GT/s吞吐率要求下最佳的SoC性能

PCIe 5.0和CXL 2.0使得32 GT/s 实现高效运行

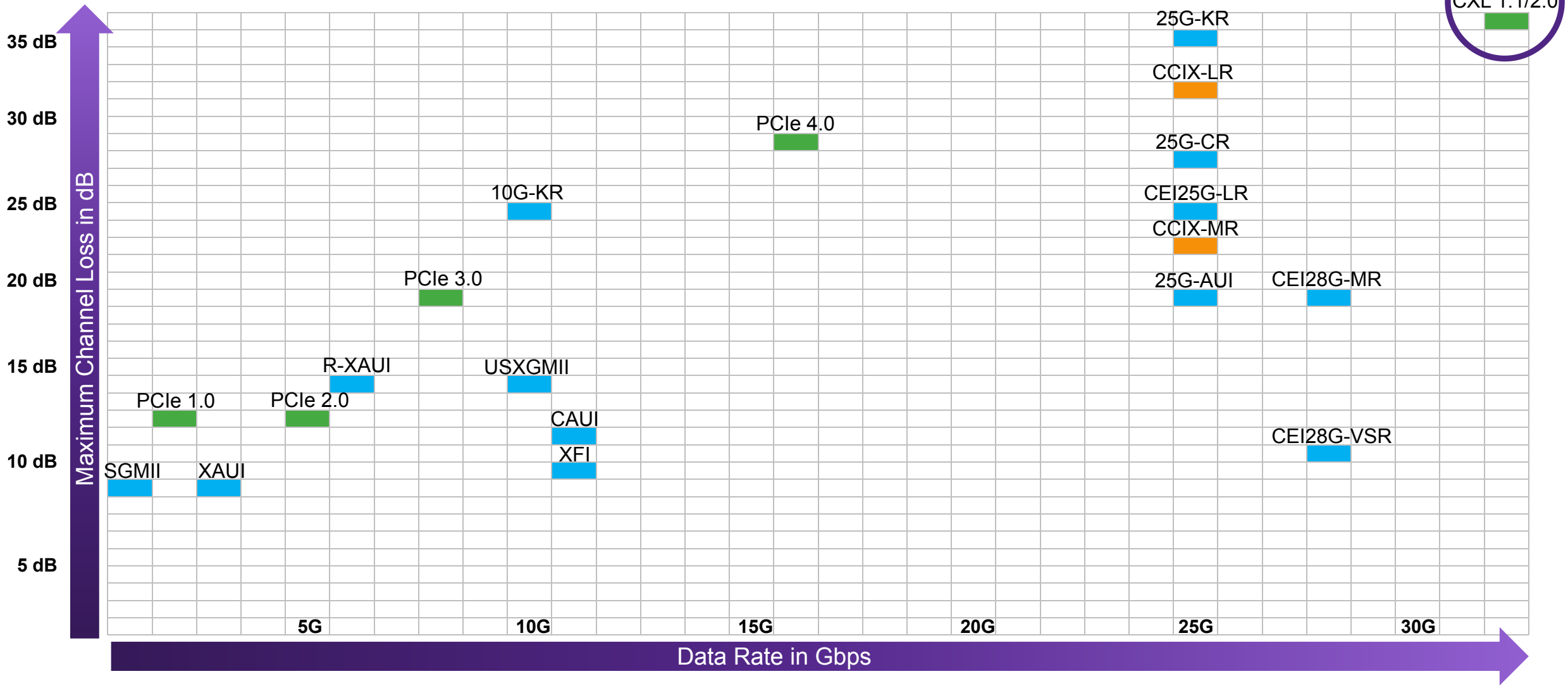
- 管理信号损伤
  - 物理层设计
  - 物理层和控制器集成
  - 封装设计
  - IBIS-AMI系统仿真
- 选择正确的控制器架构和配置
  - 可能影响性能的功能
  - 配置选项：有效载荷大小和标签
- 优化1GHz时序收敛
- 1ns UI带来最低延迟
- 针对特定的应用场景，充分利用CXL的低延迟和缓存一致性的性能

实现最高性能的32G收发器设计

关于信道的设计

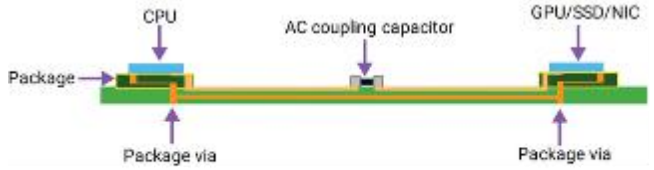
数据速率与信道损耗

PCIe 5.0和CXL必须处理最难的NRZ信道

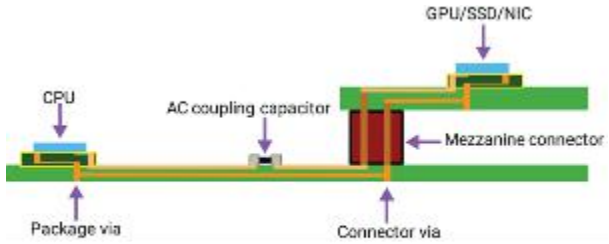


样本信道在32GT/s速率下面临的挑战

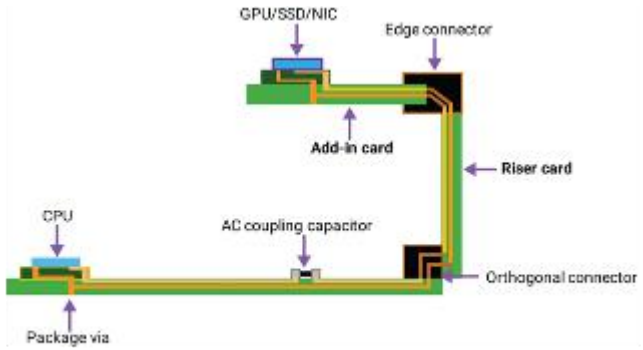
具有一系列损耗与纹波(Channel Losses & Ripple)要求的信道



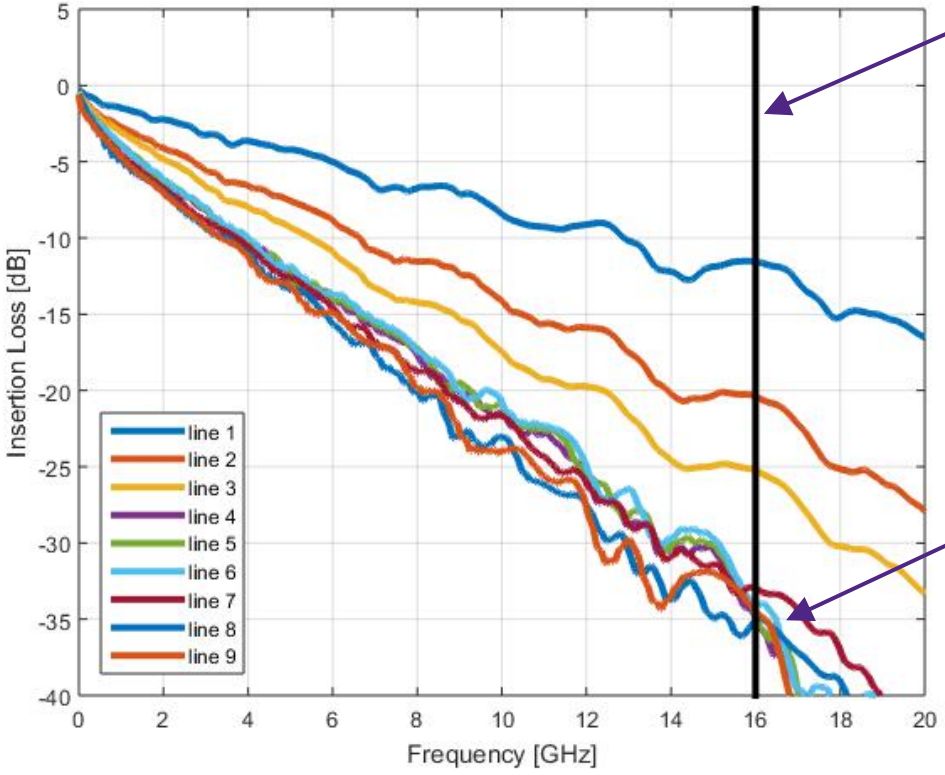
Chip-to-chip interface, the simplest channel with no connector



Channel with one mezzanine connector



Channel with two connectors using a riser card and an add-in card

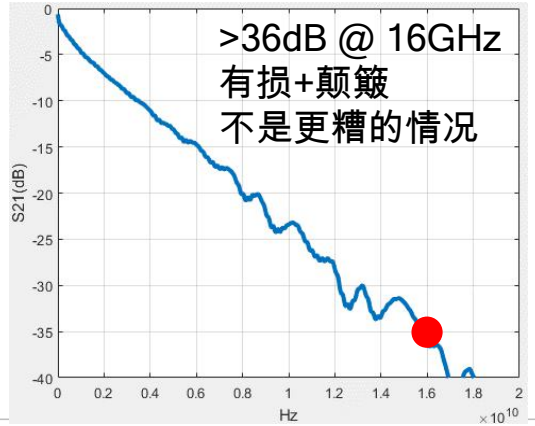


32 GT/s数据速率
(16 GHz 奈奎斯特)

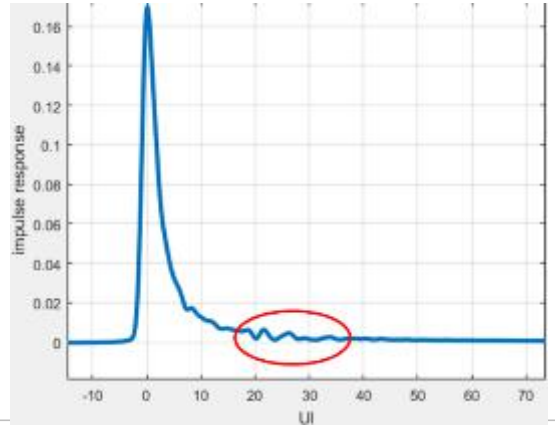
可靠的32G链路需要
在最差信道内工作

PCIe 5.0/CXL PHY性能要求

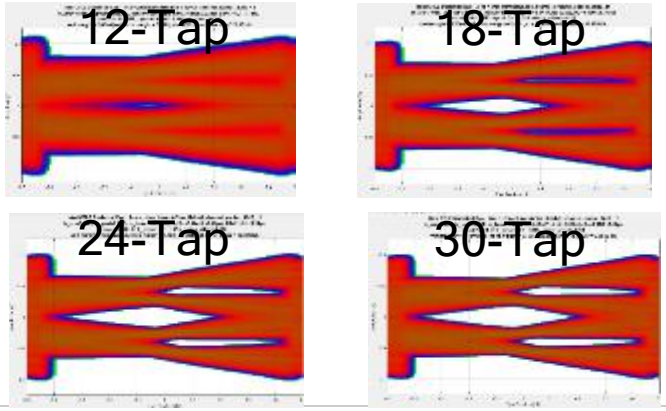
用于开发板材料的样品S21模型



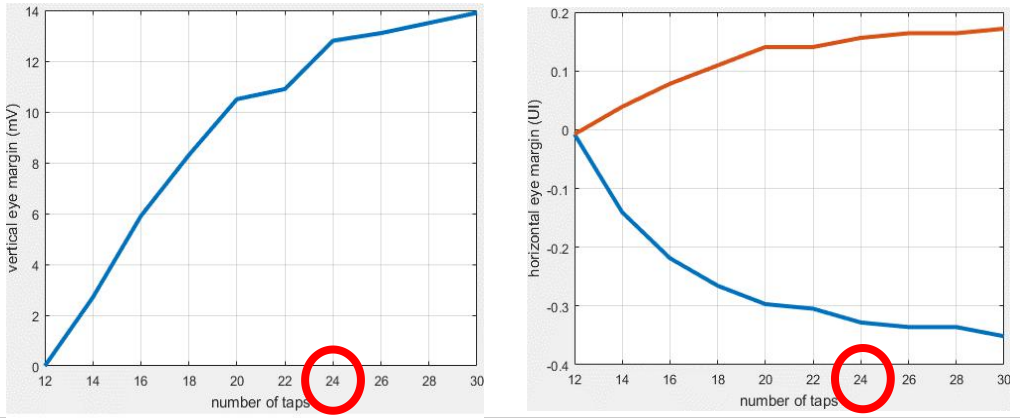
超过30的脉冲响应位会产生影响



增加DFE Tap对Open Eye的影响



Taps vs. 垂直/水平 Eye Opening

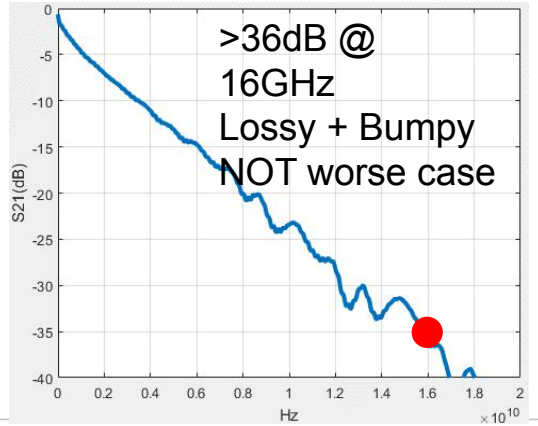


这是不是说需要一个24tap 的DFE?最好使用固定和浮动Tap的组合。

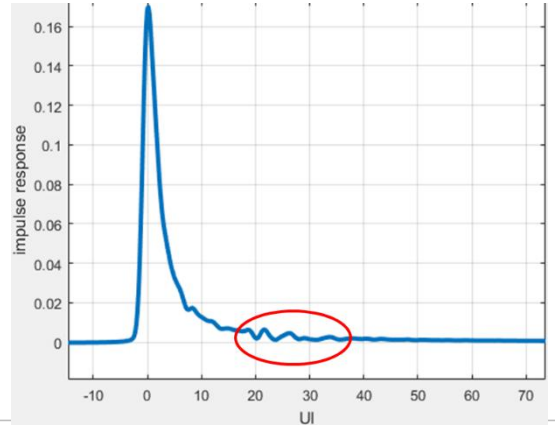
PCIe 5.0/CXL对PHY的性能要求

需要一个平衡的方法来优化PPA

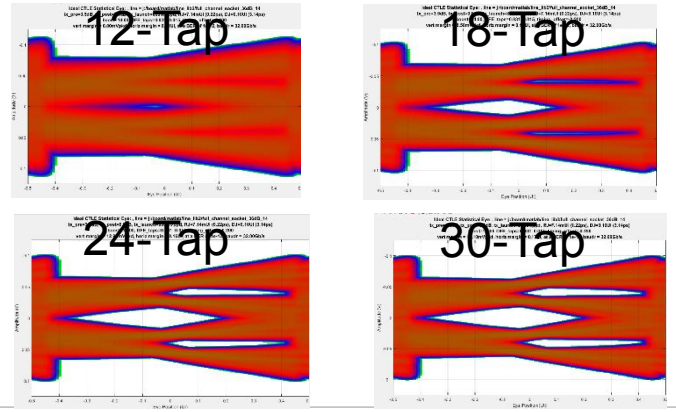
Sample S21 Model for Board Material



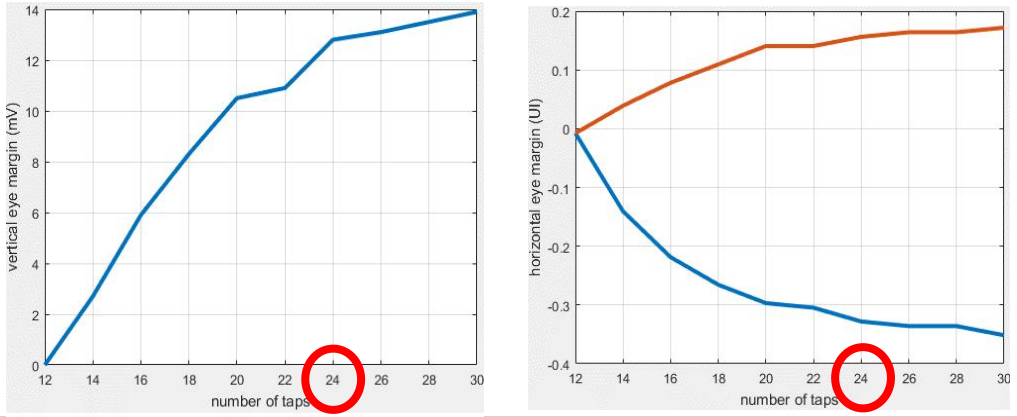
Impulse Response Bits Beyond 30 Have an Impact



Impact of Increasing DFE Taps to Open Eye



Taps vs. Vertical/Horizontal Eye Opening



这是不是说需要一个24Tap 的DFE?最好使用固定和浮动Tap的组合。

控制器的优化

控制器性能优化

实现最低延迟和最高吞吐量

ASIC架构设计

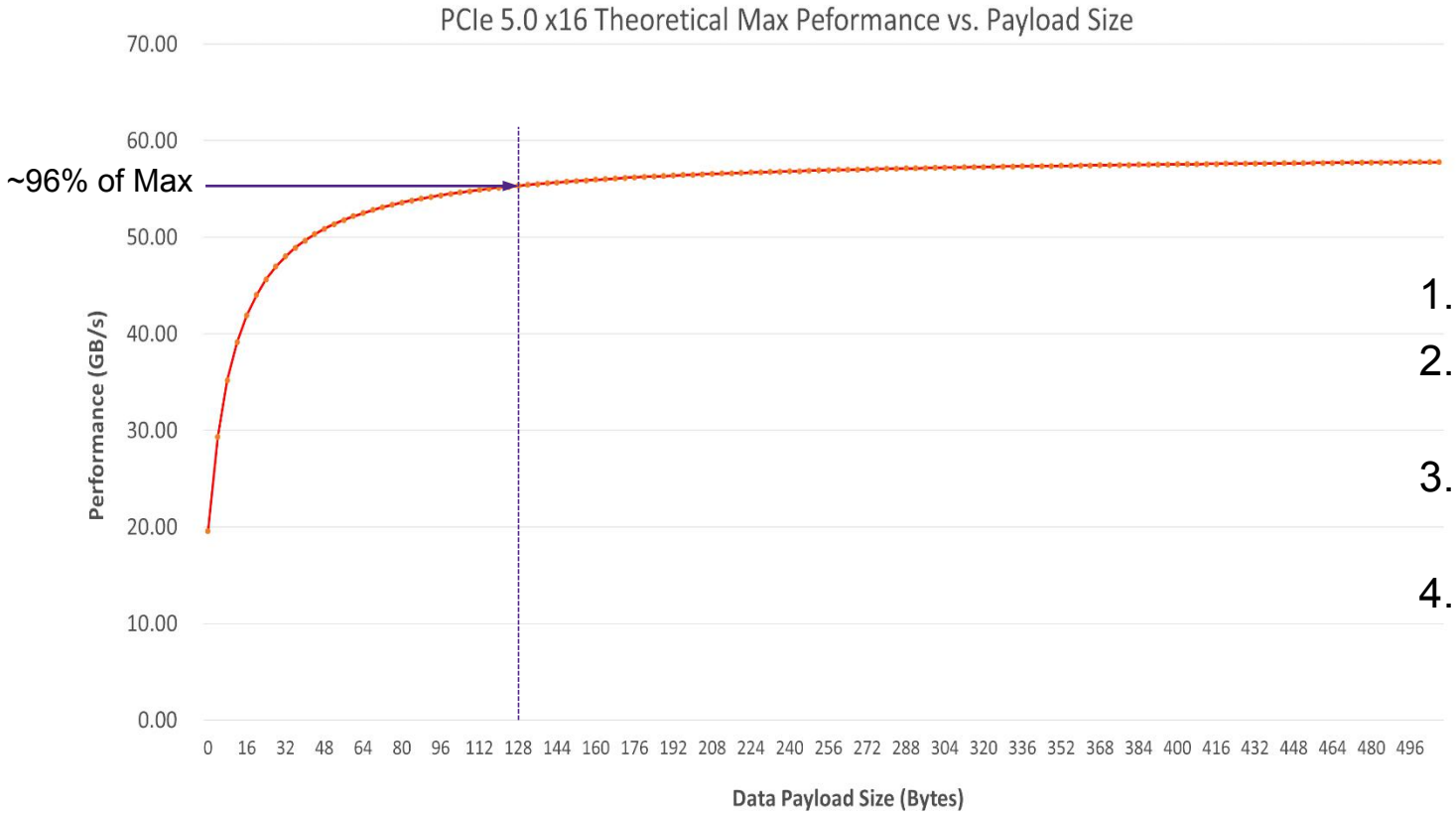
- 最小流水线级数
  - 最低延迟
  - 接近理论最大吞吐量
- 优化的ASIC架构
  - 丰富的，可选择的HPC功能集
- FPGA实现仅限于原型设计

FPGA和ASIC架构设计

- 添加流水线级数以简化计时
  - 延迟增加
- 吞吐量降低
- 能够简化设计的其他架构功能
  - 简化功能集
- 可以启用基于FPGA的产品

注意负载的大小以获得最佳性能

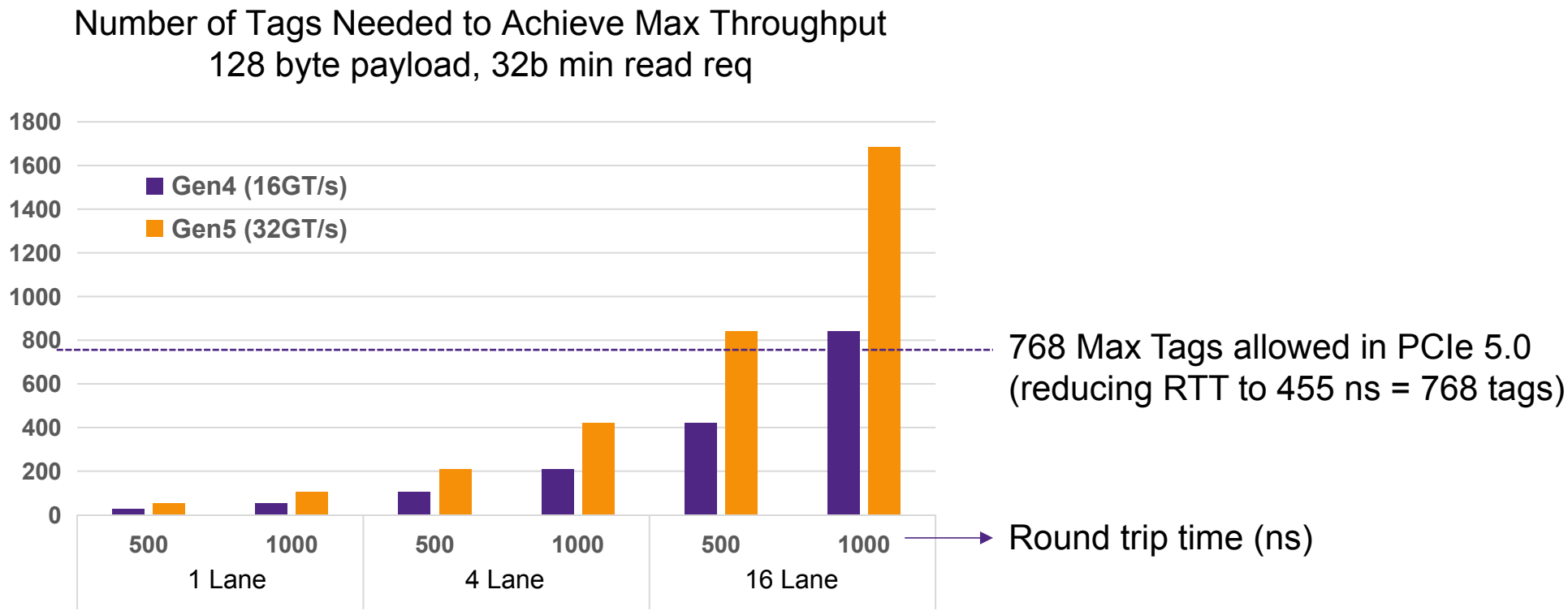
16 Lanes @ 32GT/s Per Lane模式下: 小的有效载荷效率非常低



1. Max = 32GT/s x 16 lanes = 512e9 T/s
2. Change to Gbytes/Sec = 512,000,000,000 /(8*1024^3) = 59.605GB/s
3. Account for Header, LCRC, Seq & Framing= 59.605 * ~0.985 (@ 508 Bytes) = 58.688GB/s
4. Account for 128b/130b Encoding = 58.688 *(128/130) = 57.785GB/s

控制器的配置问题

设置正确的标签数量对于PCIe 5.0（32GT/s）更为重要



标签数量会限制未完成未发布读取请求。
通道越多，往返时延就越高，速度越快=保持链路饱和需要更多标签

优化核心配置以提高性能

示例：正确设置标签数量会带来性能的巨大提升

PCIe 5.0 Link Bandwidth with 2 Tags									
Direction	Link	Total Transfer Size (bytes)	Average Transfer Size (bytes)	Traffic breakdown (%)	Data B/W (GB/s)	Link Ideal (GB/s)	Link Loss (%)	Core Ideal (GB/s)	Core Loss (%)
RX	GEN5 x16	32768	256.00	UPDATE_FC_NP: 2.740; Replays: 0 CPL_D/3DW/NO_ECRC: 94.541; ACK: 2.719	2.872	41.444	93.070	41.444	93.070
TX	GEN5 x16	0	0	Replays: 0	0	0	0	0	0

PCIe 5.0 Link Bandwidth with 32 Tags									
Direction	Link	Total Transfer Size (bytes)	Average Transfer Size (bytes)	Traffic breakdown (%)	Data B/W (GB/s)	Link Ideal (GB/s)	Link Loss (%)	Core Ideal (GB/s)	Core Loss (%)
RX	GEN5 x16	32768	256.00	UPDATE_FC_NP: 0.604; Replays: 0 CPL_D/3DW/NO_ECRC: 98.792; ACK: 0.604	38.058	43.307	12.120	43.307	12.120
TX	GEN5 x16	0	0	Replays: 0	0	0	0	0	0

PCIe 5.0 Link Bandwidth with 128 Tags									
Direction	Link	Total Transfer Size (bytes)	Average Transfer Size (bytes)	Traffic breakdown (%)	Data B/W (GB/s)	Link Ideal (GB/s)	Link Loss (%)	Core Ideal (GB/s)	Core Loss (%)
RX	GEN5 x16	32768	256.00	UPDATE_FC_NP: 0.382 ; Replays: 0 CPL_D/3DW/NO_ECRC: 99.147; ACK: 0.471	55.351	55.351	-0.000	55.351	-0.000
TX	GEN5 x16	0	0	Replays: 0	0	0	0	0	0

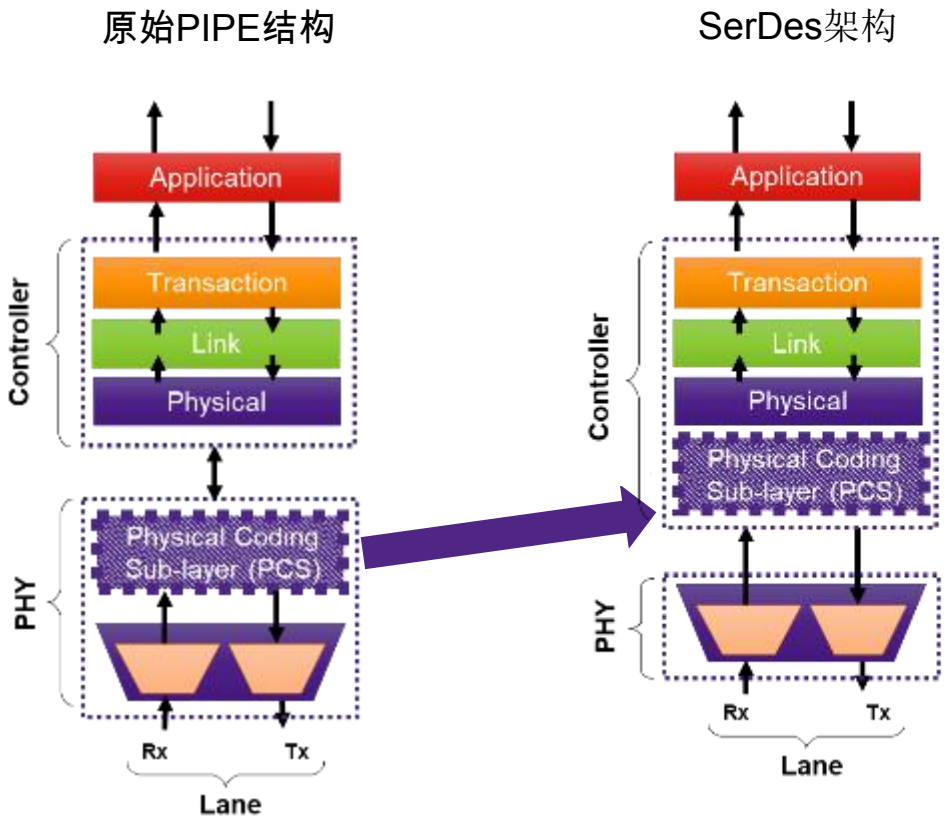
Data from Synopsys coreConsultant tool

PHY与控制器的集成

PIPE 接口更新-PIPE 5.1.1/5.2.1

PHY与控制器关键接口之间的影响

- PIPE 5.1.1/5.2.1
  - 低接口引脚数（LPC）：映射PIPE信号到消息总线寄存器
  - 增加了对PCIe 5.0（ 32GT/s），聚合IO和DisplayPort的支持
  - 增加了对“ SerDes”架构的支持（10位，20位，40位，80位选项）
  - 支持64位PIPE （80位）（适用于SerDes架构）



许多PCIe客户首选的解决方案是具有原始PIPE架构的PIPE 5.1.1 / 5.2.1

在PIPE接口模式下不高于1GHz的时序收敛要求

已证明512b 模式对CXL和PCIe 5.0至关重要

- HPC设计需要最大带宽
  - 16通道的PCIe 5.0
  - 16通道的CXL 1.1或CXL 2.0
- 1GHz下的时序收敛意味着必须使用32位模式(32GT/s ÷ 1GHz = 32b)
- 32b PIPE x 16 lanes =512b datapath interface
- 需要经过量产验证的512位控制器架构

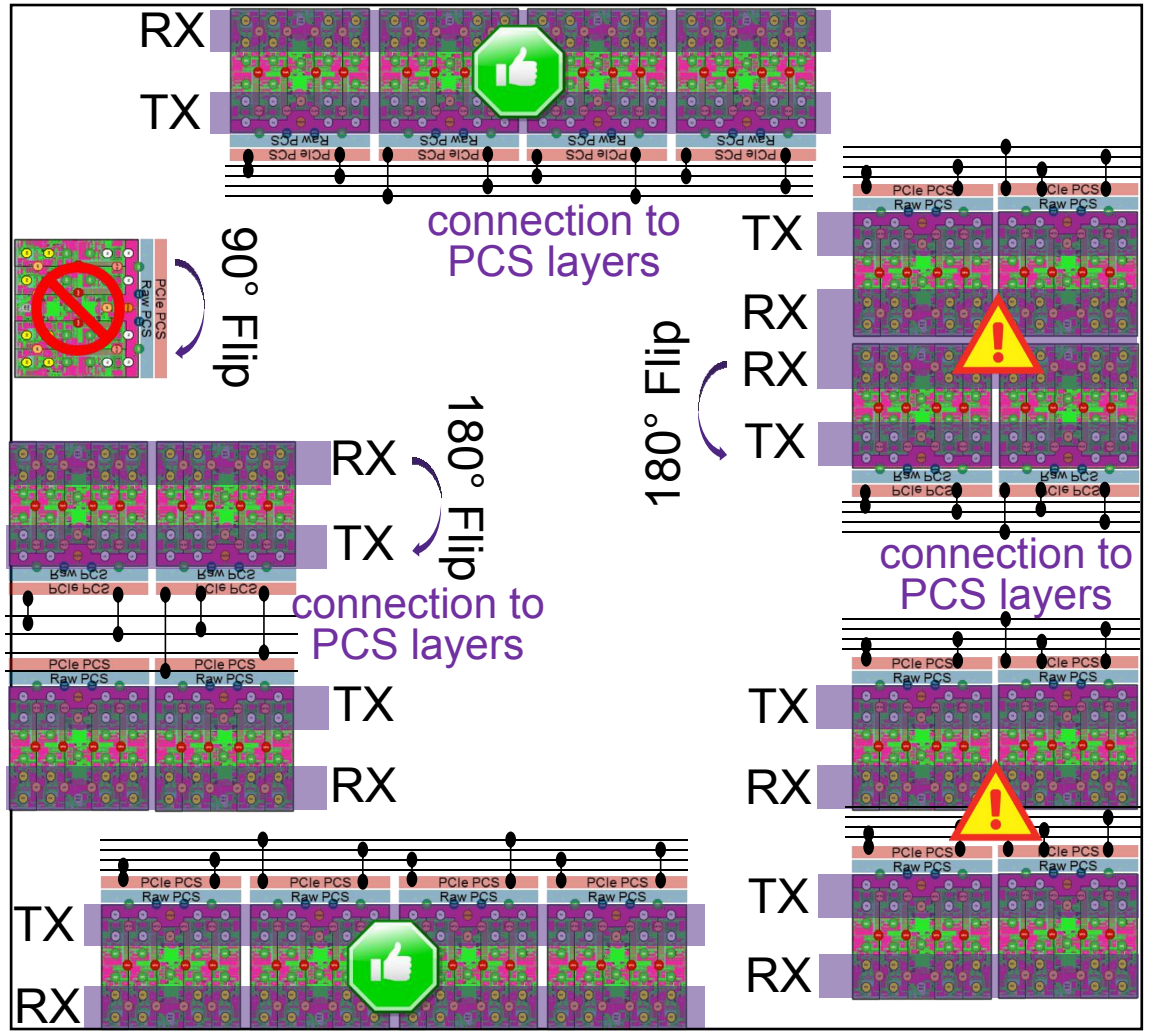
	PCIe 5.0 或 CXL @ 32GT/s		PCIe 4.0 16GT/s
	x16	X8 (对CXL而言非典型)	x16
总带宽	512GT/s	256GT/s	256GT/s
512b时序收敛	**1GHz**	**500MHz**	**500MHz**
256b时序收敛	**2GHz**	**1GHz**	**1GHz**

非常困难！

32G PHY 的物理位置很重要

影响控制器的实现并简化时序收敛

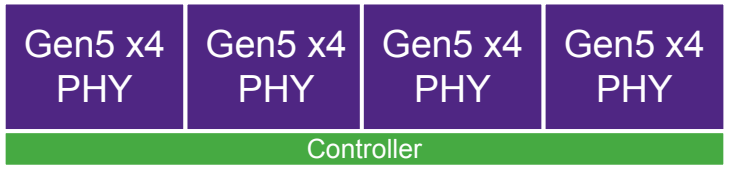
- 优化硬核间距准则可用于支持所有边缘上的布局
- ESD电路嵌入在PHY内，靠近PHY的I/O Bump
- 多边形必须为N-S（不允许90°翻转）；可以将IP放置在PCS-PCS首选方向的E-W边缘上：
- 硬核也可以支持在两排摆放，以方便SoC进行高密度设计



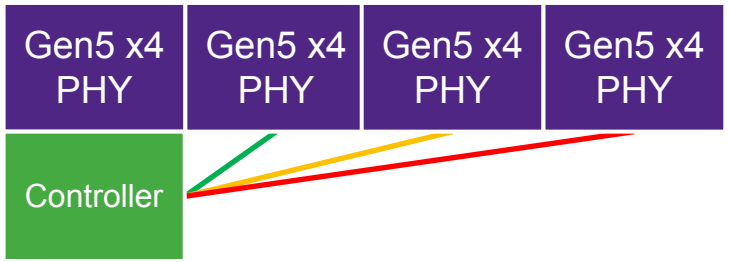
达到1GHz时序收敛的注意事项

宽链接宽度和1GHz时钟频率组合需要在P＆R中格外注意

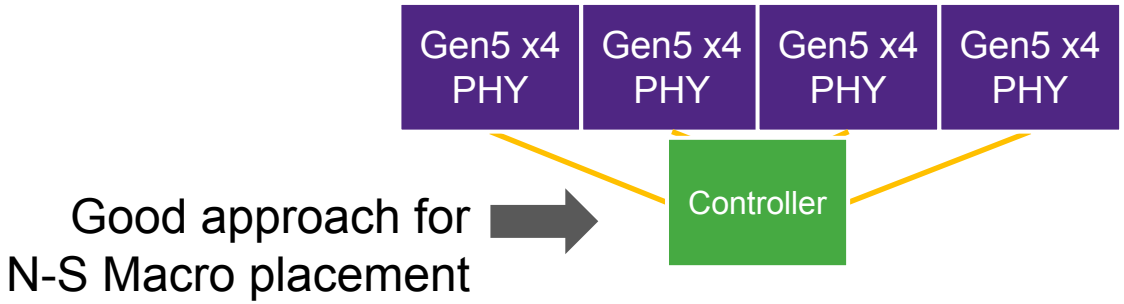
First inclination to spread controller across wide PHY may prevent needed clustering of logic

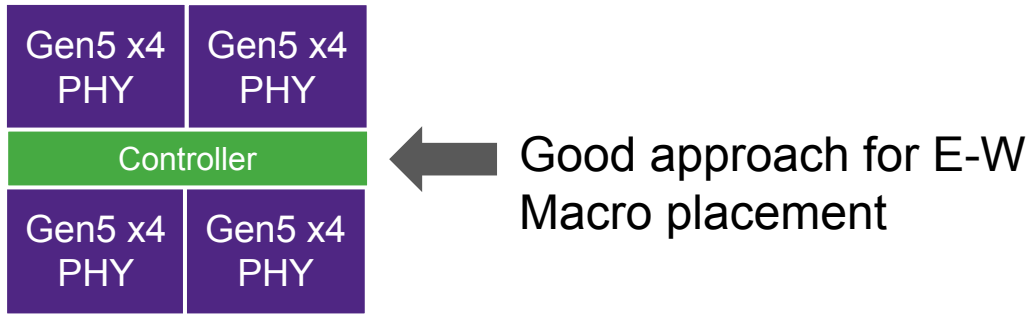


Clumping controller at one end or the other creates extremely long runs for some lanes



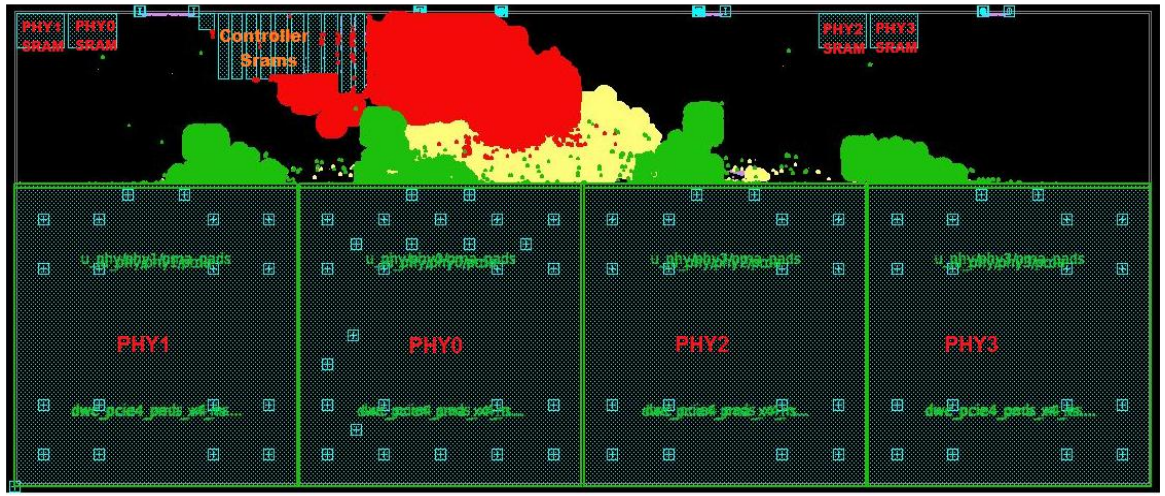
Best results come from balanced approaches, depending on N-S, E-W placement





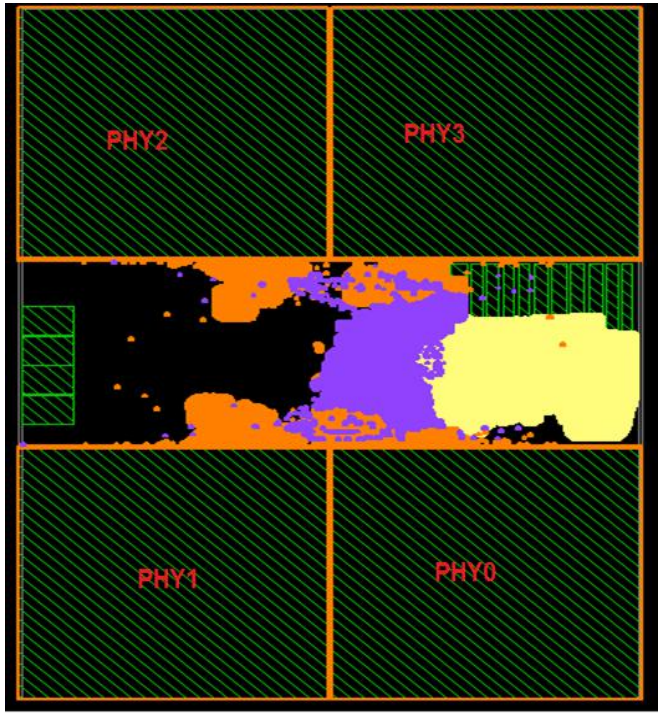
利用Synopsys解决方案成功完成1GHz时序收敛的结果

1GHz Timing for N-S Macro Placement



Red = Controller hierarchy
Yellow = UPCS hierarchy
Green = Raw PCS hierarchy

1GHz Timing for E-W Macro Placement



Yellow = Controller hierarchy
Purple = UPCS hierarchy
Orange = Raw PCS hierarchy

在32GT/s速度下的信号完整性

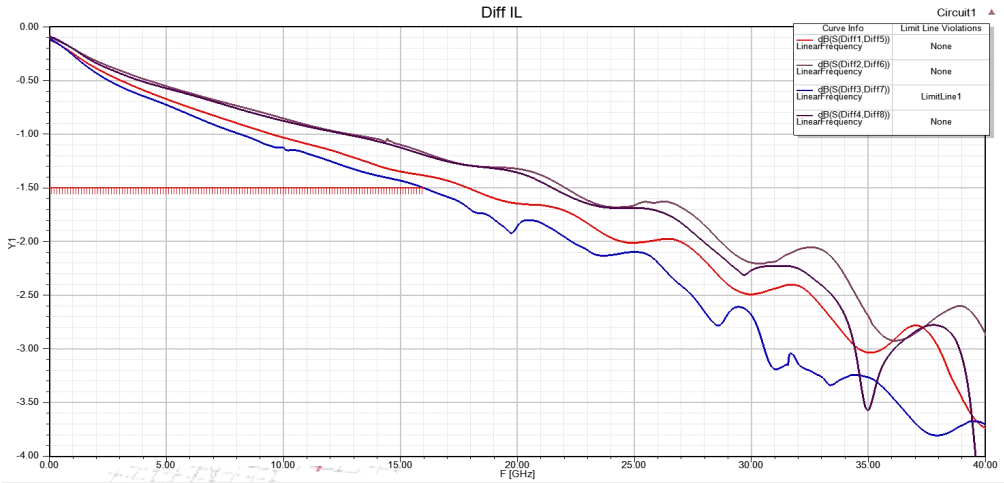
减少连接中断、重试、恢复行程

32GT/s速度要求下的封装设计新的设想

规范需要扩展到16GHz，满足新的规格是越来越困难

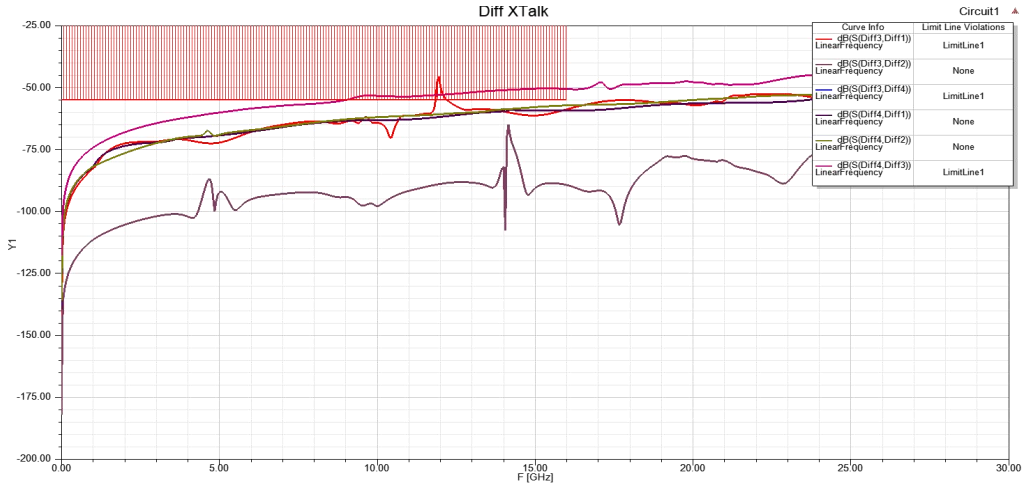
Meeting differential insertion loss specification

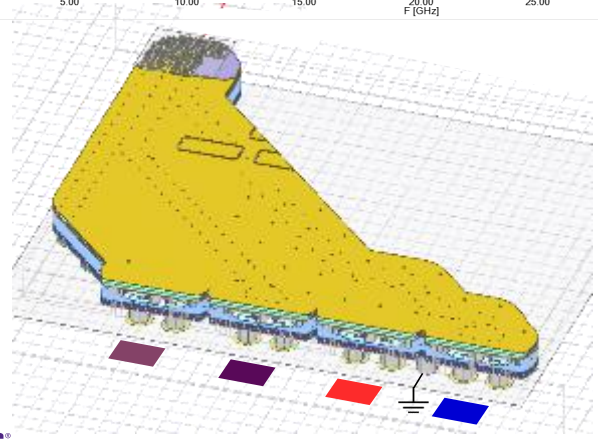
Example specification for 32G: -1.5 dB up to 16GHz



Meeting differential crosstalk specification

Example specification for 32G: -55 dB up 16 GHz





Package design and modeling

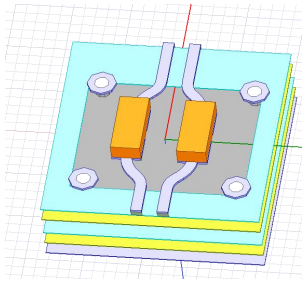
- Trace length and routing must be carefully managed within the specific package form factor

反射和串扰在32GT/s速率下问题更为突出

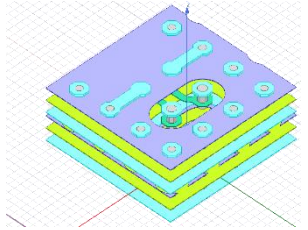
影响封装、连接器和电路板设计

- 整个互连的反射对接收器均衡提出了挑战
  - 互连不同点上的不连续性(例如:Package Bump，BGA焊球，连接器，通孔，隔直电容等）
    - 考虑整个系统和元件来匹配阻抗
  - 在通孔区域中的不当布线会增加相邻通道之间的串扰
    - 尽量保持走线的最大间距，即使在拥挤的通孔区域也是如此

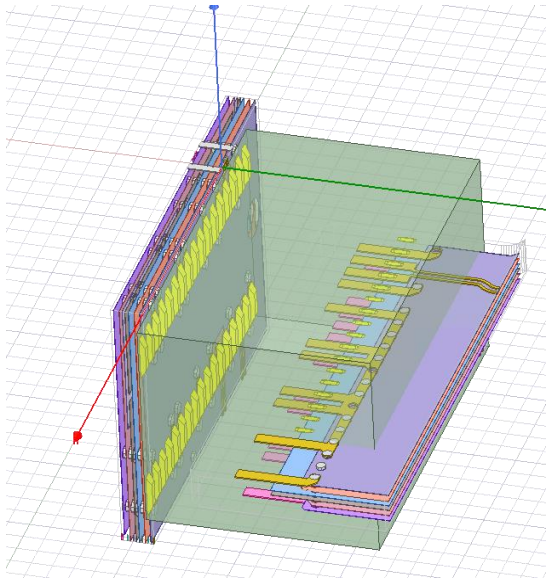
DC Blocking Cap



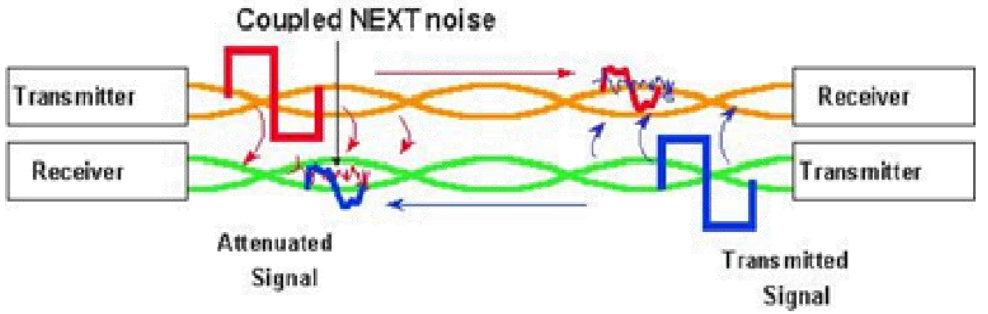
VIAs



Connector



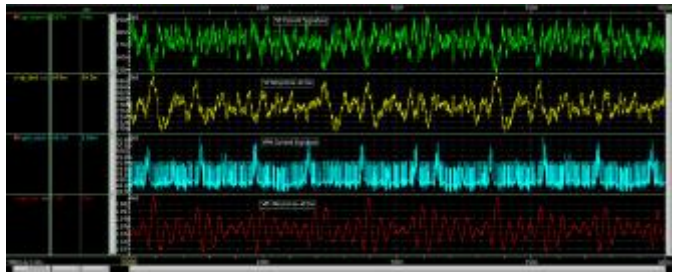
- 相邻TX和RX通道之间的近端串扰（Near-end-cross-talk）在32GT/s时更严重
  - 做好TX和RX布线，以避免串扰



管理32GT/s模式下关键的电源传输

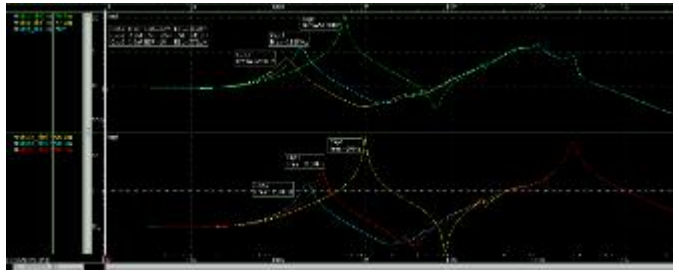
电源完整性

- 随着速度的提高……
  - 幅值和带宽对电源电流需求在增加
  - 保持芯片上稳定电源电压的基本挑战依然存在
- 32G PHY IP的片上电源必须在一系列操作条件下保持稳定
  - 任务模式操作-所有通道以全数据速率传输
  - 一个通道中的电源状态变化（模式变化），可能会在连续传输模式下为其他通道产生过多的di/dt事件
- 电源传输分析必须涵盖整个电源网络，包括PCB、封装和芯片组件
- 来自其它模内电路的噪声必须与高速SerDes接口有效隔离，反之亦然



任务模式下的交流纹波仿真

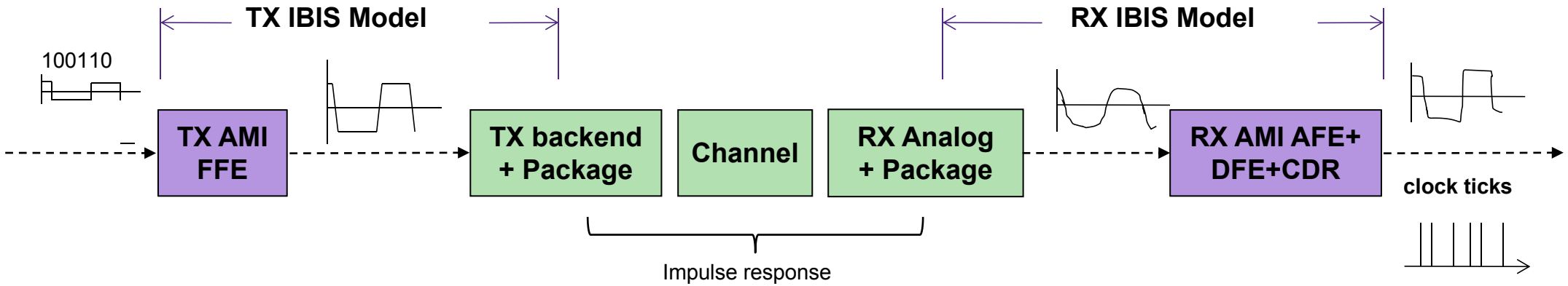
验证模式改变对其他通道的影响

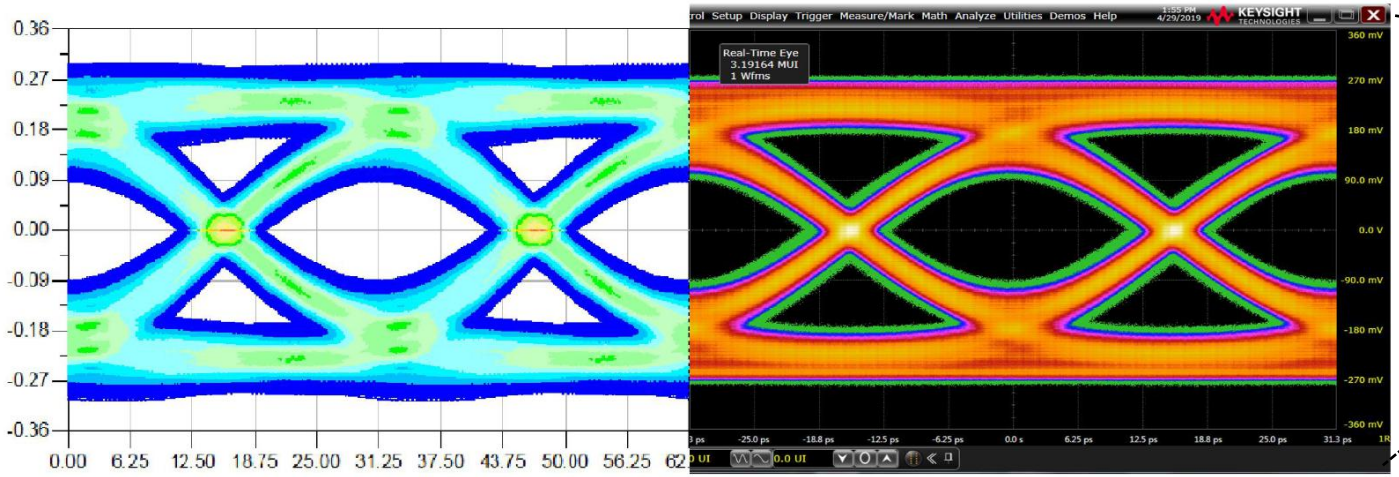


板上滤波器频率响应元件

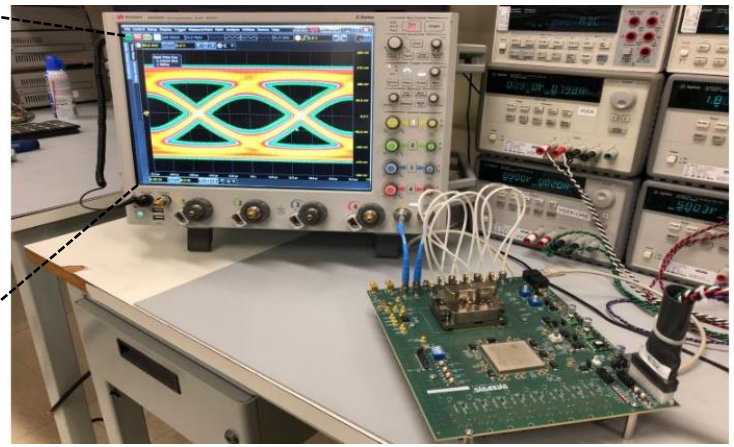
IBIS AMI 模型与PCI5.0与CXL（32GT/s）需求

正确的使用可以带来非常准确的结果，并支持完整的系统模拟





IBIS AMI Simulation



Measured Data

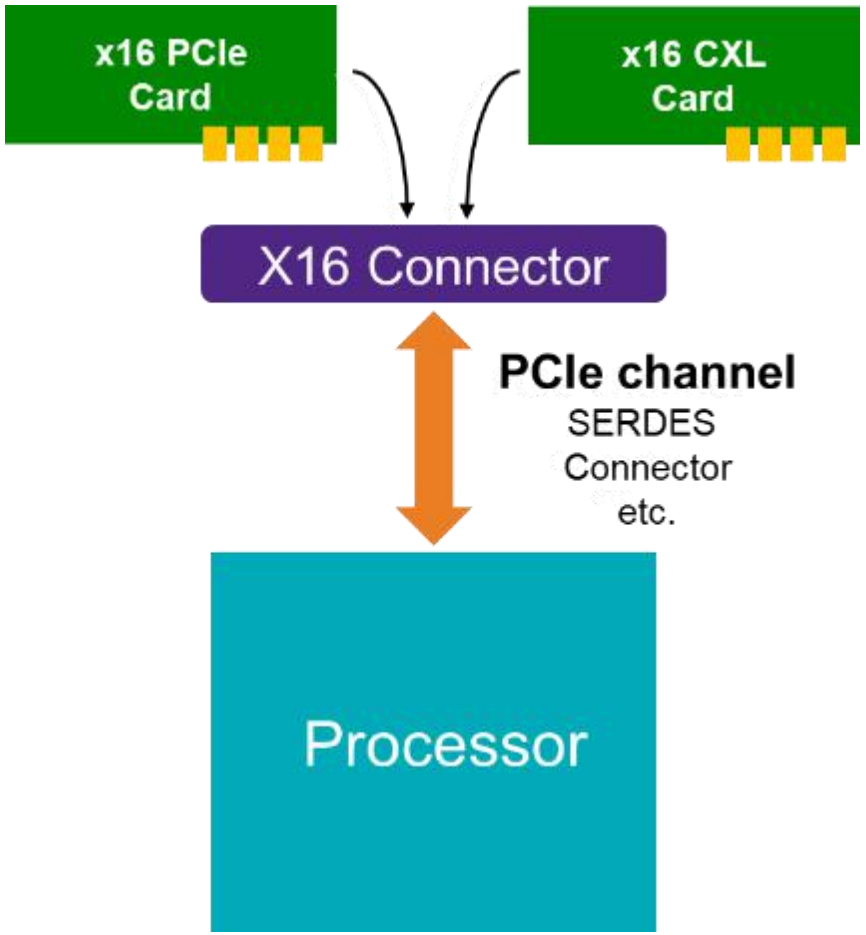
关于CXL

革命性的延迟规范需要定制的解决方案

Compute Express Link (CXL)

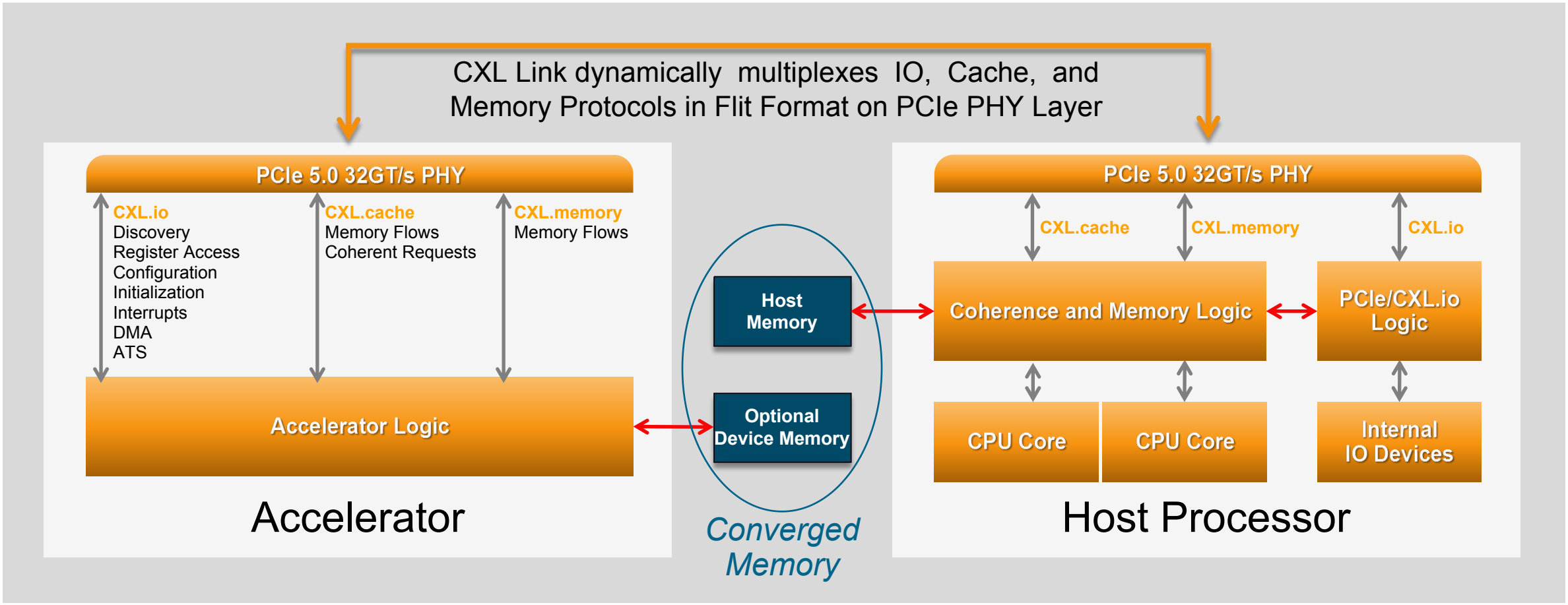
CXL是什么?

- CXL是一种协议，运行在标准PCIe物理层上
- CXL可以自动协商到标准PCIe协议或备用CXL协议
- CXL允许CPU和加速器访问彼此的内存，提供3种协议：
  - I/O（CXLio）基本上在CXL模式下承载的PCIe流量（最大4KB） 必选*
  - Cache（CXLcache）用于缓存主机内存的设备的一致协议 可选*
  - Memory（CXL.mem）用于设备内存的小块（最大64B）低延迟协议 可选*
- CXL使用PCIe 5.0 电气和接口特性，和PCIe可以在相同的接口上操作
- 从2021/2022年开始，Intel Sapphire Rapids CPU支持PCIe 5.0和CXL



CXL内存共享

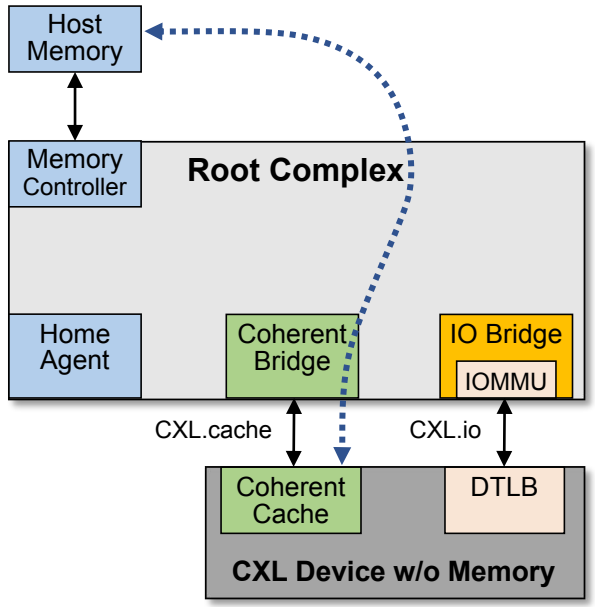
CXL支持主机和设备之间的聚合内存



一种跨越处理器和设备的单一、通用的内存地址空间

CXL为3种不同的设备类型定义了独特的应用程序

Type 1 Device



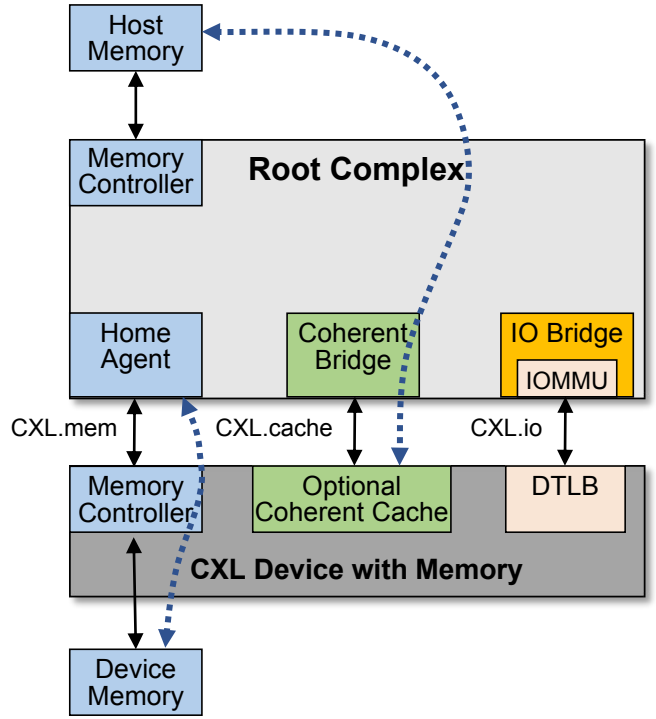
Caching Devices and Accelerators:

Accelerators, Smart NICs
(NIC Atomics & shared address space)

CXL.io + CXL.cache

Device coherently accesses Host memory

Type 2 Device



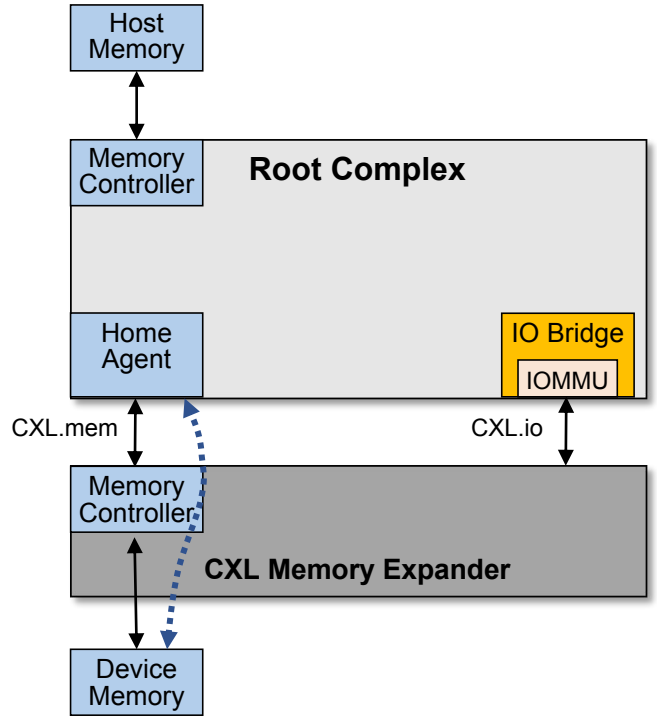
Device with attached memory

GPU, Dense computation

CXL.io + CXL.cache + CXL.mem

Device coherently accesses Host memory
Host accesses Device memory

Type 3 Device



Memory expander devices

Memory BW and/or capacity expansion
Persistent memory

CXL.io + CXL.mem

Host accesses and manages attached Device memory

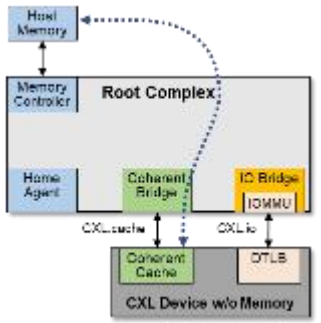
CXL应用正在增长

- 加速器接口（全部利用卸载和数据移动）
  - AI 和 HPC
  - 数据库的计算与存储
  - 网络加速器
- 内存接口（扩展到CPU附加内存之外）
  - 存储类内存和DRAM（加载存储内存）
- CXL外部的桥接和网关，用于计算访问大型结构中的组件
  - 更广泛的内存结构
  - 分类计算

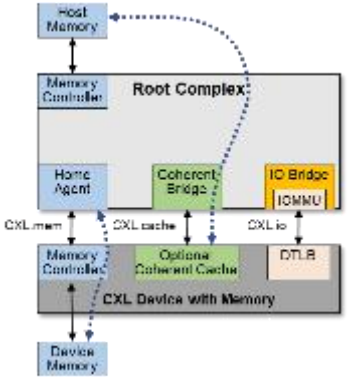
桥接Gen-Z

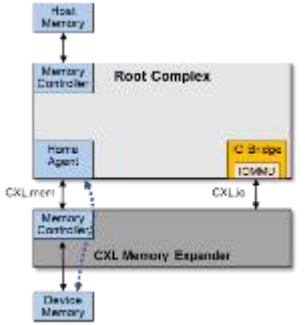
- 新的支持CXL的内存形式

1型



2型





3型

谁需要CXL？

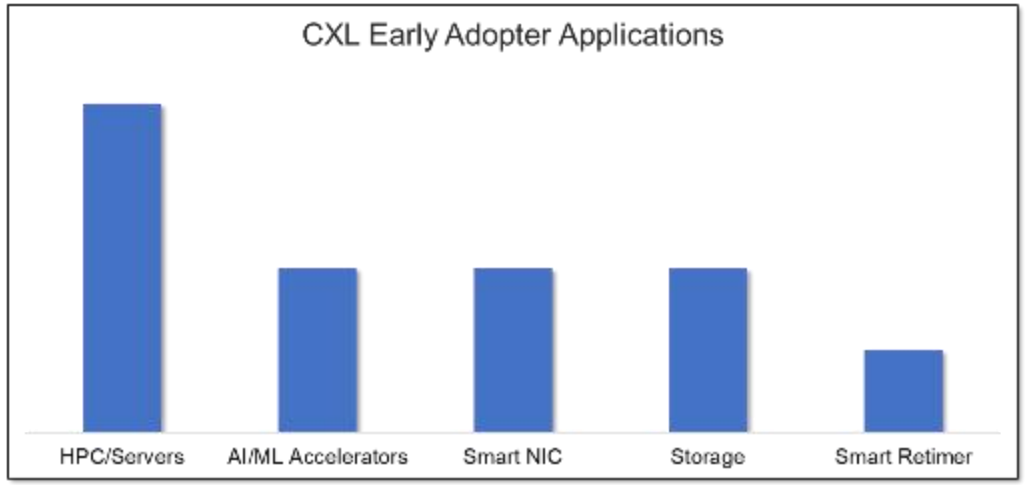
不是所有人…

需要

- 计算应用
  - 人工智能处理器
    - 汽车高级驾驶辅助系统
  - 机器学习系统
    - 面部识别引擎
  - 其他协处理器缓存系统内存
    - 智能网络信息中心
- 内存扩展设备
  - 线性存储器
    - 字节可寻址（vs块）SSD后续程序
  - 非易失性存储器控制器
    - NVDIMM、MRAM等
  - 专用内存扩展
    - GDDR?、计算存储?,其他未来应用?

不需要

- 数据移动应用程序
  - 传统存储控制器
  - 传统网络接口
- CXL培训材料还讨论了何时不使用CXL

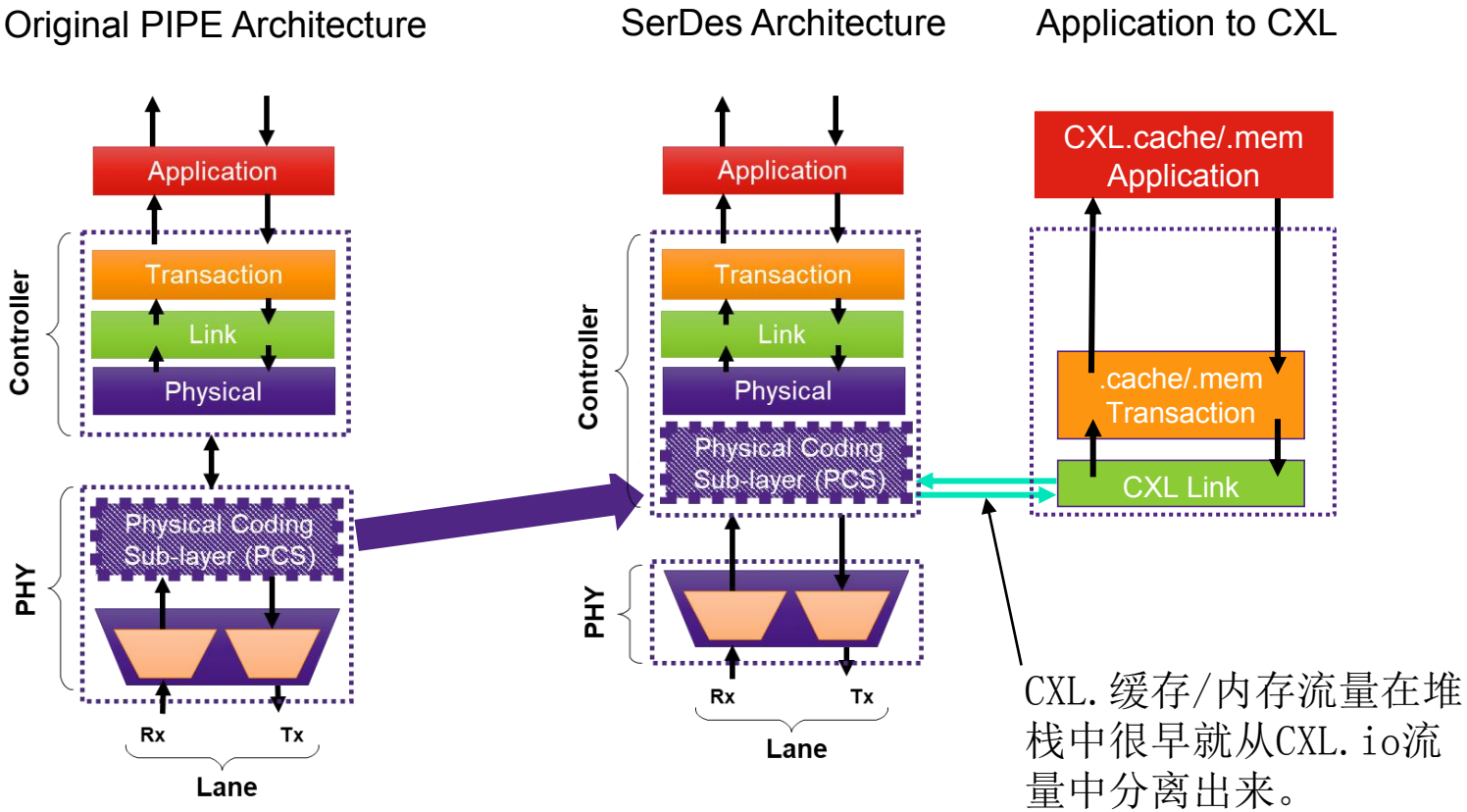


Based on actual Synopsys customer data

重新审视用于CXL的PIPE接口

正确处理接口对于实现低延迟至关重要

- SerDes 架构是关键
  - 将PCS功能从PHY移动到控制器
  - 现在CXL绕过了这个延迟（10、20、40、80位选项）
- CXL控制器设计应该是.cache和.mem的基础设计
  - 简化这些路径
  - 面向40ns往返延迟规范

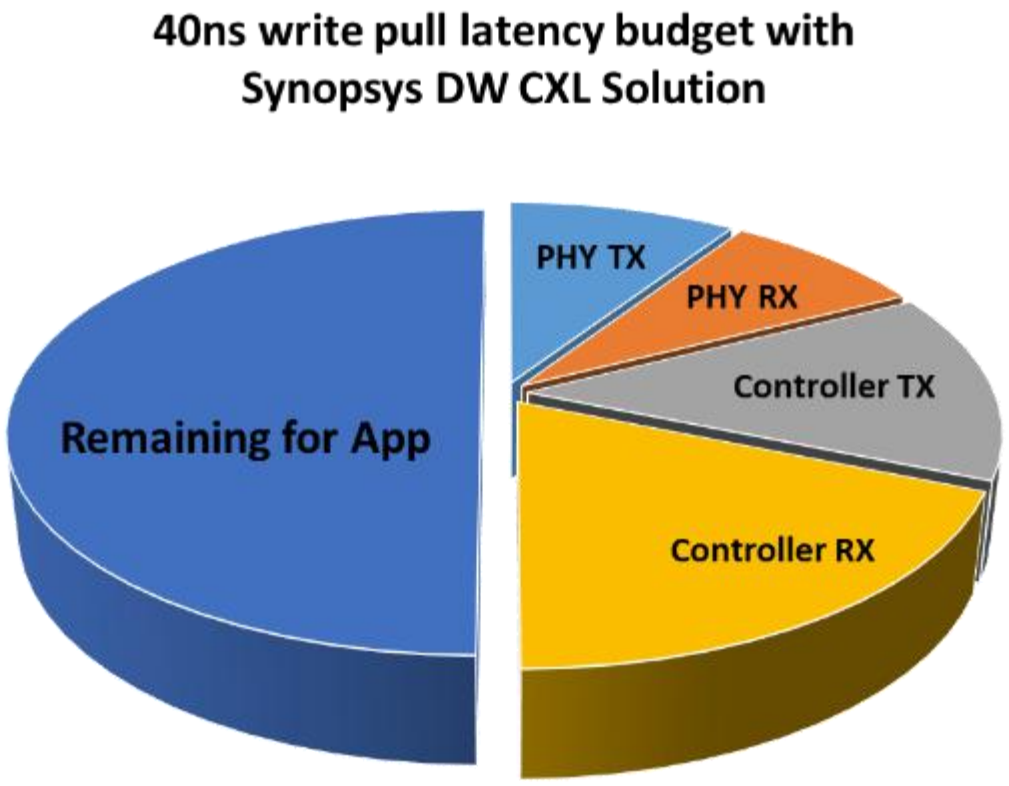


在CXL解决方案中实现低延迟

Synopsys的DesignWare CXL控制器+PHY解决方案

Synopsys延迟-优化解决方案

- 采用PIPE5.1.1/5.2.2 SerDes架构
- 优化了32G PHY，减少了PMA和PC的延迟
- 为您的应用程序提供总预算的50%



用于PCIe 5.0和CXL的DesignWare CXL IP方案

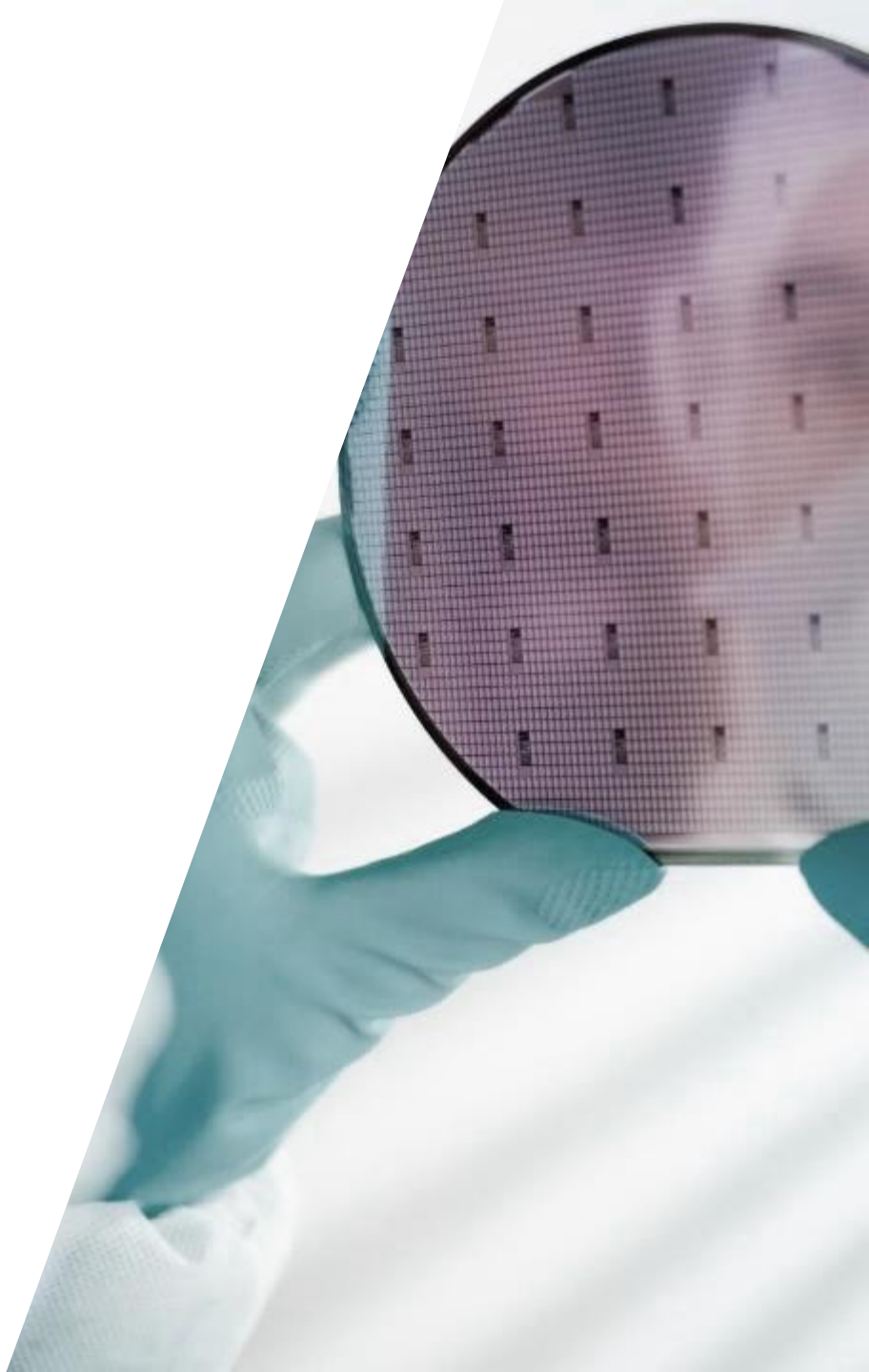
面向高性能计算SoCs

Synopsys PCIe和CXL优势概述

	Synopsys	其他	评论
PCIe Gen4、Gen5和CXL的领导者，风险最低	✓	x	• Synopsys有更多的客户，出带（tapeout）和客户硅 • 200多个Gen4许可证，100多个Gen5许可证，数十个CXL许可证 • *Synopsys在PCIe集成商列表中以22个条目领先*
完全针对SOC进行优化，具有更好的性能	✓	x	• 最高吞吐量（超过理论值的98%），延迟最低 • 延迟，吞吐量优势均已通过客户验证，无论大小负载
先进的RAS-DES简化了调试和软件开发	✓	x	• Synopsys拥有最广泛的RAS-DES功能，可覆盖多达64000个信号 • 在推出新标准（如Gen5和CXL）时尤为重要
在许多客户配置中，1GHz时序收敛得到认证	✓	x	• 在32GT/s数据速率和32bPIPE条件下，对Gen5和CXL至关重要 • 经过多种工艺技术验证，包括28nm，16nm和7nm
经过生产验证的512b架构	✓	x	• PCIe 5.0和CXL x16链路必备 • 自2016年以来，Synopsys 512b已通过Si验证并获得许可> 60次=最低风险
在您的办公桌上探索架构权衡	✓	x	• PCIe 5.0或CXL控制器+coreConsultant=快速、轻松优化 • 调整配置参数并在几分钟内看到模拟结果
配置拦截提供了独特的功能	✓	x	• 允许客户的应用程序拦截配置读写 • 提升应用意识；可以启用硅后错误修复
可扩展IOV提供灵活的SRIOV支持	✓	x	• 支持高达64K的虚拟功能；VFs可以分布在任何PFs上 • 外部存储器和控制器内部的实现
灵活的接口支持	✓	x	• Synopsys本机接口提供最高性能 • 用于Arm一致互连的Arm CXS和AXI接口（CMN-XXX） • 支持Arm SMMU的DTI-ATS、LTI和MSI-GIC接口
32G PHY，经过众多代工厂验证，强大耐	✓	x	• 32G PHY满足所有PCIe 5.0规范，并提供广泛的铸造支持 • 生产过程中提供客服，提供Char报告
丰富的PCIe互操作和硬件演示	✓	x	• 业界最丰富的互操作计划确保了与所有供应商的完美合作 • Synopsys完整的PCIe 5.0系统硬件演示在多个会议上展示

总结

- 现在是开启PCIe 5.0或CXL 1.1/2.0 32G 项目的好时机
- 32G设计必须处理标准中最困难的NRZ通道
  - 32G的插耗达到36dB及以上
  - 32G的信道不稳定，有很多不连续性
- 需要精心设计收发器
  - 必须仔细考虑AFE+CTLE+DFE，并共同努力缓解问题
  - 寻找先进的DFE来处理问题，直至达到24位
- 32G的PHY和控制器集成需要更仔细的规划
  - PIPE5.2.1兼容性
  - 平面图和时序收敛
  - 使用完整的解决方案（PHY+控制器）以获得最佳结果
- 注意控制器的配置会严重影响在32GT / s的性能
- SIPI分析，封装和板级设计至关重要
- 将IBIS-AMI模型集成到您的系统仿真计划中
- 选择合适的IP和设计合作伙伴以确保成功



SYNOPSYS®
Silicon to Software™

Thank You