面向数据中心的高性能 SDN交换技术

深圳市风云实业有限公司

2016年10月

1. 云数据中心网络的发展趋势
2. Nebula Flex Connect 技术介绍

云数据中心网络的发展趋势

应用层

物理层

智能化的应用软件
具备可编程能力的硬件
自动化的配置和管理

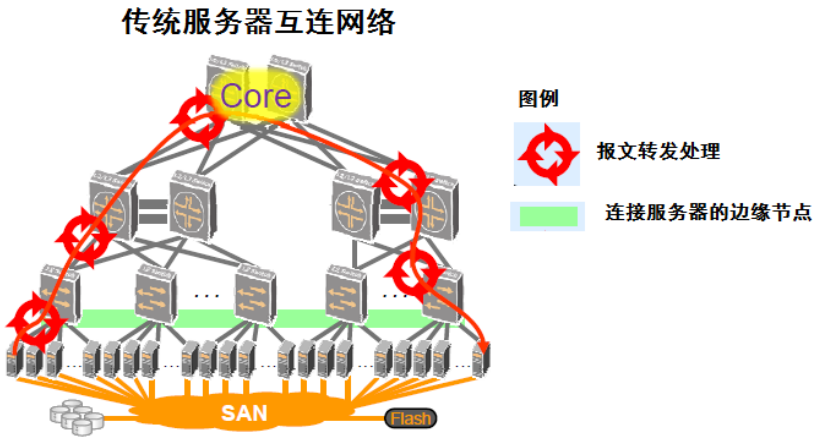
来自客户的疑问

- 敏捷性
- 安全性
- 自动化
- 移动性
- 解耦合

- 数据交换 → 大规模、高密度、低延迟、高吞吐、集约化、弹性化

传统通信网络互连服务器带来如下问题：

- 通过交换机级联构建网络
- 各个服务器的性能与其网络位置相关，服务器节点之间带宽与通信路径强相关，带宽受限
- 每一步交换都要进行报文转发处理，解析、查表、封装报文，增大服务器节点之间跳步，增加了通信延迟大，通信性能随跳步数降低
- 静态负载均衡，不能感知网络状态，避免拥塞，不可预知的通信热点，造成路径拥塞，加剧延迟的抖动，链路利用率低，网络性能在重负载下急剧下降
- 交换机的配置复杂繁琐，难以维护管理
- 仅支持以太网、FC或者Infiniband 单一网络接口



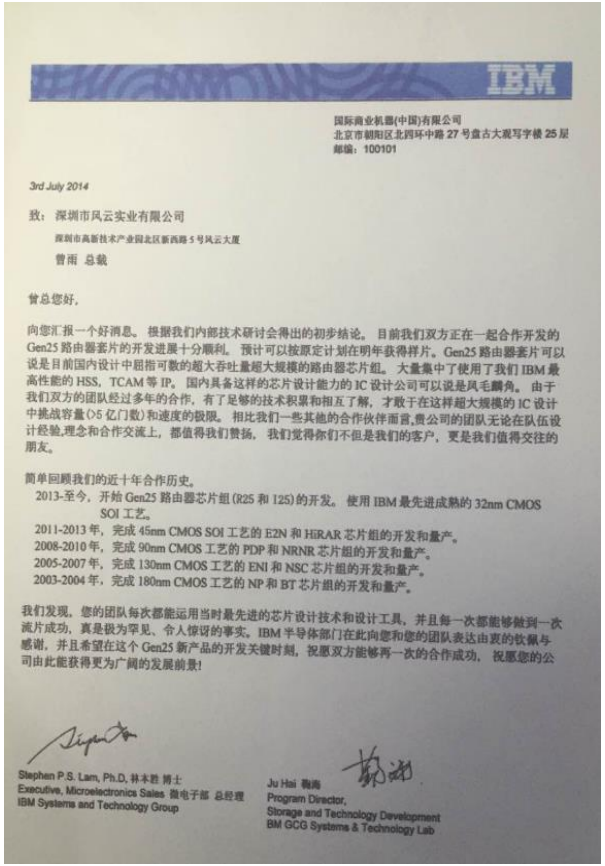
1. 云数据中心网络的发展趋势
2. Nebula Flex Connect 技术介绍

Nebula互联方案概述

- **定位：**面向大数据与云计算中心市场，消除传统互连网络的弊端，设计研发基于标准以太网和PCIE协议接口的**高性能、多协议、跨平台**的高性能互连系统核心技术，支持**SDN应用场景**，高效的支持**虚拟化和融合**网络。
- **核心技术：**高带宽、低延迟、高可用、可编程的互连芯片技术为核心。

  两款核心的高端互连芯片：通用可编程处理芯片Korath和高密度交换芯片Collector。

  配套的开发和管理软件IOMC，以及相关原型系统的机框、机架、板卡的参考设计。
- **知识产权情况：**风云公司和IBM拥有所有相关知识产权，相互授权对方不受约束的研发、生产、销售和改进技术。



IBM

国际商业机器(中国)有限公司
北京市朝阳区北四环中路 27 号盘古大观写字楼 25 层
邮编：100101

3rd July 2014

致：深圳市风云实业有限公司
深圳市高新技术产业园北区新西路 5 号风云大厦
曾雨 总裁

曾总您好，

向您汇报一个好消息。根据我们内部技术研讨会得出的初步结论，目前我们双方正在一起合作开发的Gen25路由器套片的开发进展十分顺利。预计可以按原定计划在明年获得样片。Gen25路由器套片可以说是目前国内设计中启指可数的超大吞吐量超大规模的路由器芯片组。大量集中了使用了我们 IBM 最高性能的 HSS，TCAM 等 IP。国内具备这样的芯片设计能力的 IC 设计公司可以说是凤毛麟角。由于我们双方的团队经过多年的合作，有了足够的技术积累和相互了解，才敢于在这样超大规模的 IC 设计中挑战容量(5 亿门数)和速度的极限。相比我们一些其他的合作伙伴而言,贵公司的团队无论在队伍设计经验,理念和合作交流上，都值得我们赞扬，我们觉得你们不但是我们的客户，更是我们值得交往的朋友。

简单回顾我们的近十年合作历史。

2013-至今，开始 Gen25 路由器芯片组(R2S 和 12S)的开发。使用 IBM 最先进成熟的 32nm CMOS SOI 工艺。
2011-2013 年，完成 45nm CMOS SOI 工艺的 E2N 和 H2RAR 芯片组的开发和量产。
2008-2010 年，完成 90nm CMOS 工艺的 PDP 和 NRNR 芯片组的开发和量产。
2005-2007 年，完成 130nm CMOS 工艺的 ENI 和 NSC 芯片组的开发和量产。
2003-2004 年，完成 180nm CMOS 工艺的 NP 和 BT 芯片组的开发和量产。

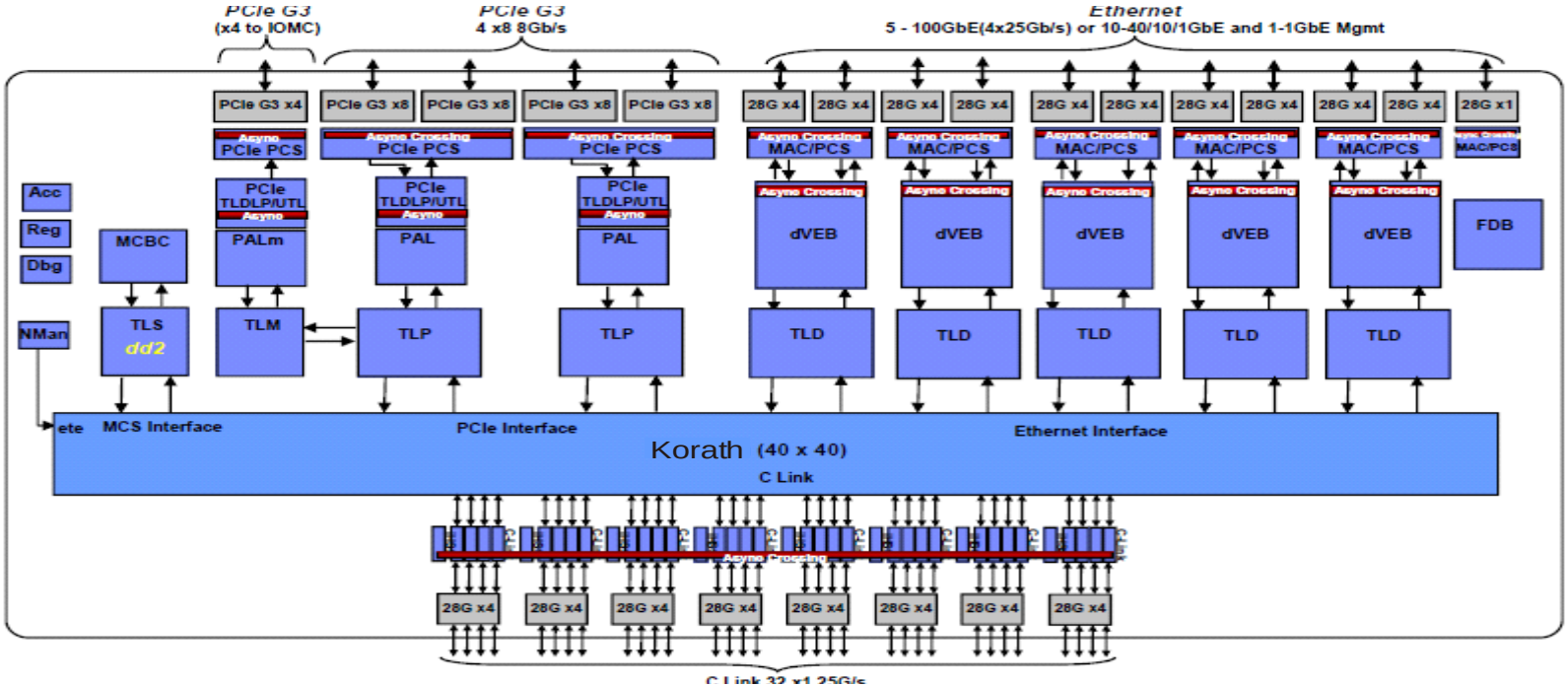
我们发现，您的团队每次都能运用当时最先进的芯片设计技术和设计工具，并且每一次都能够做到一次流片成功，真是极为罕见、令人惊讶的事实。IBM 半导体部门在此向您和您的团队表达由衷的钦佩与感谢，并且希望在这个 Gen25 新产品的开发关键时刻，祝愿双方能够再一次的合作成功，祝愿您的公司由此能获得更为广阔的发展前景!

Stephen P.S. Lam, Ph.D, 林本胜 博士
Executive, Microelectronics Sales 微电子部 总经理
IBM Systems and Technology Group

Ju Hai 鞠海
Program Director,
Storage and Technology Development
IBM GCG Systems & Technology Lab

Korath芯片架构

SDN的尝试和思考



Korath芯片技术规格

（1）性能指标

- 5路100GE 以太网接口/10路40GE/10GE、5路 PCIE GEN3接口、32路动态均衡的25Gbps C-link接口，提供共计2Tbps 带宽
- 最小转发延迟250ns、30MB 片上报文缓冲
- 支持HASH+TCAM结合的多级流表、内嵌多核可编程协处理器
- 支持136个Korath芯片互连扩展能力(同时支持680个100G 端口和544个PCIE端口)

（2）支持完整的以太网特性

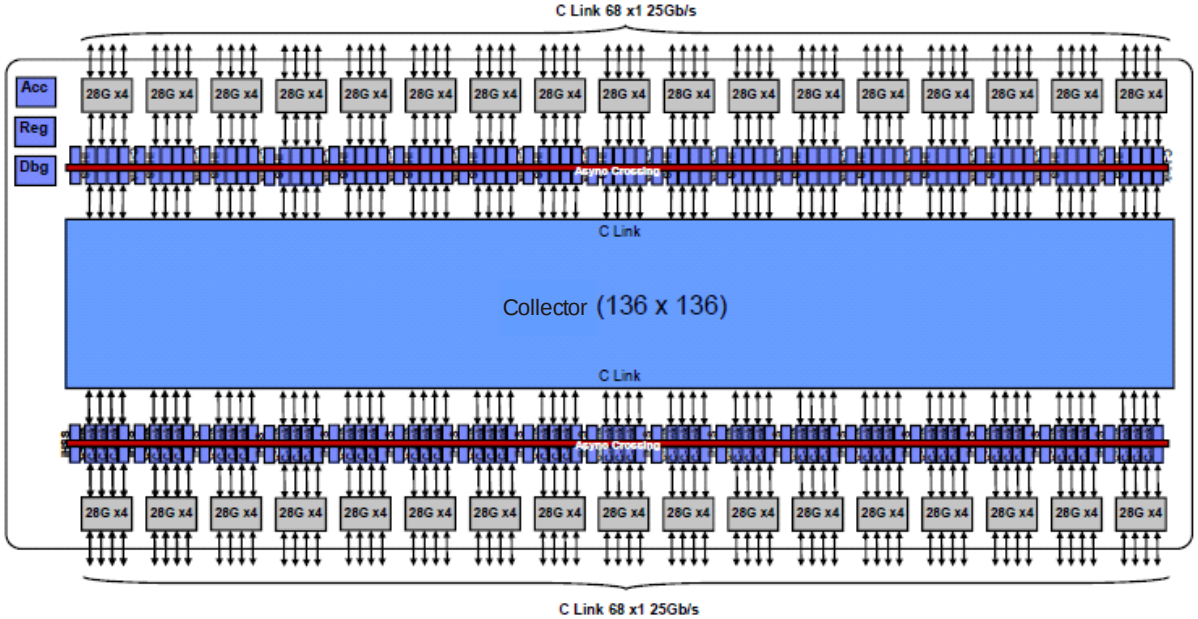
- 支持传统以太网协议
- 支持数据中心扁平交换特性
- 支持数据中心无损以太网协议

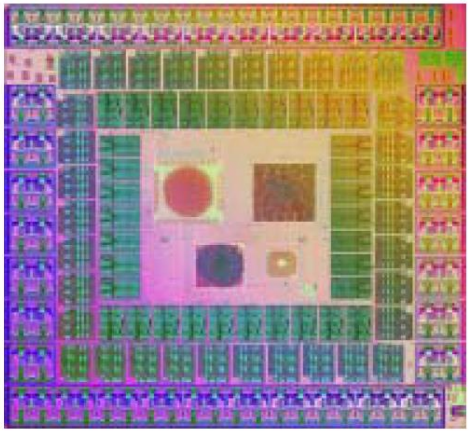
（3）丰富的PCIE交换特性

- 支持每个PCIE端口可以独立配制成上行或下行端口，上行端口接服务器的PCIE Root Port，下行端口接PCIE Adapter
- 支持灵活配置的分布式PCIE Switch，每个上行端口可以连接最大32个下行端口，每个下行MR-IOV端口支持64个上行端口
- 支持单根虚拟化（SR_IOV），支持多根虚拟化（MR_IOV),支持最大64个虚拟层次(VH)、
- 支持通过MATCH-ACTION的机制实现PCIE接口数据交换

Collector芯片架构

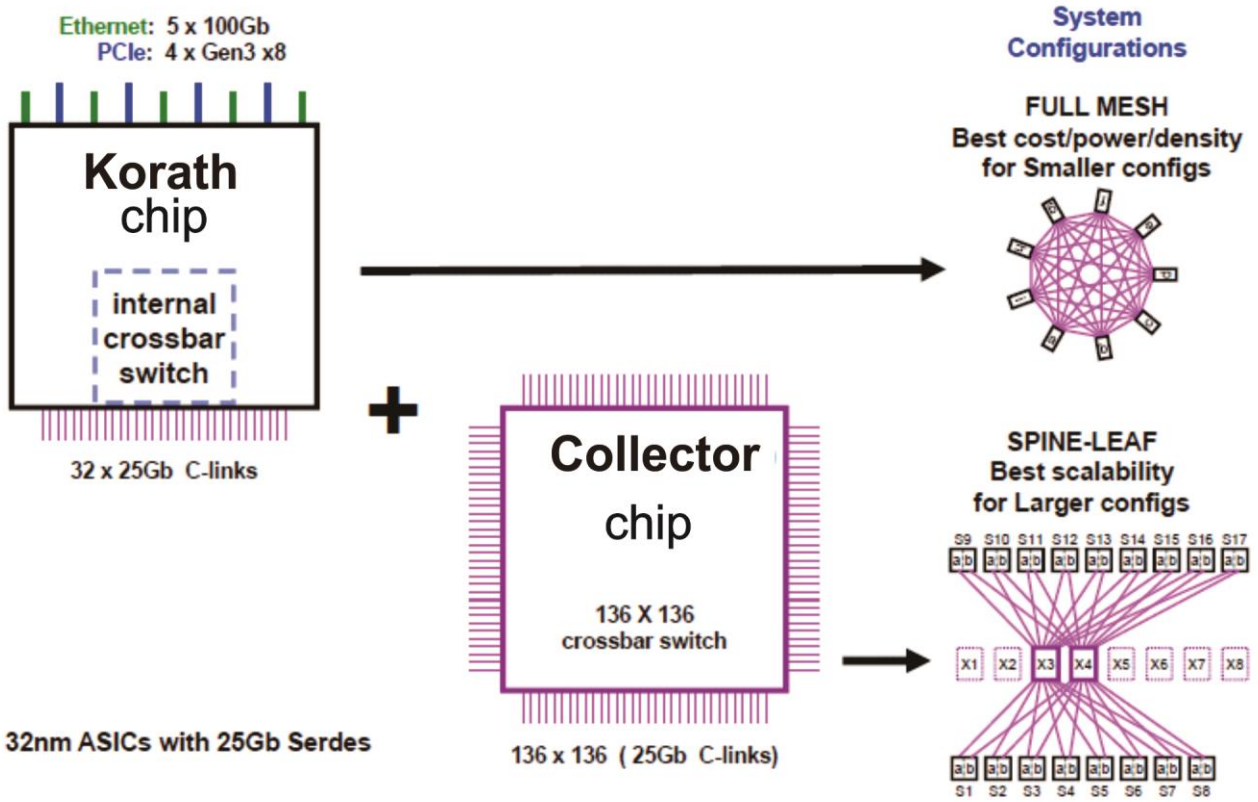
- 136个25G C-link端口，交换带宽6.8Tbps
- 无阻塞交换，交换延迟175ns
- 支持端到端可靠性（自动CRC检查与重传机制）
- 支持芯片级自适应路由、支持芯片级的多路径负载均衡和等价路由



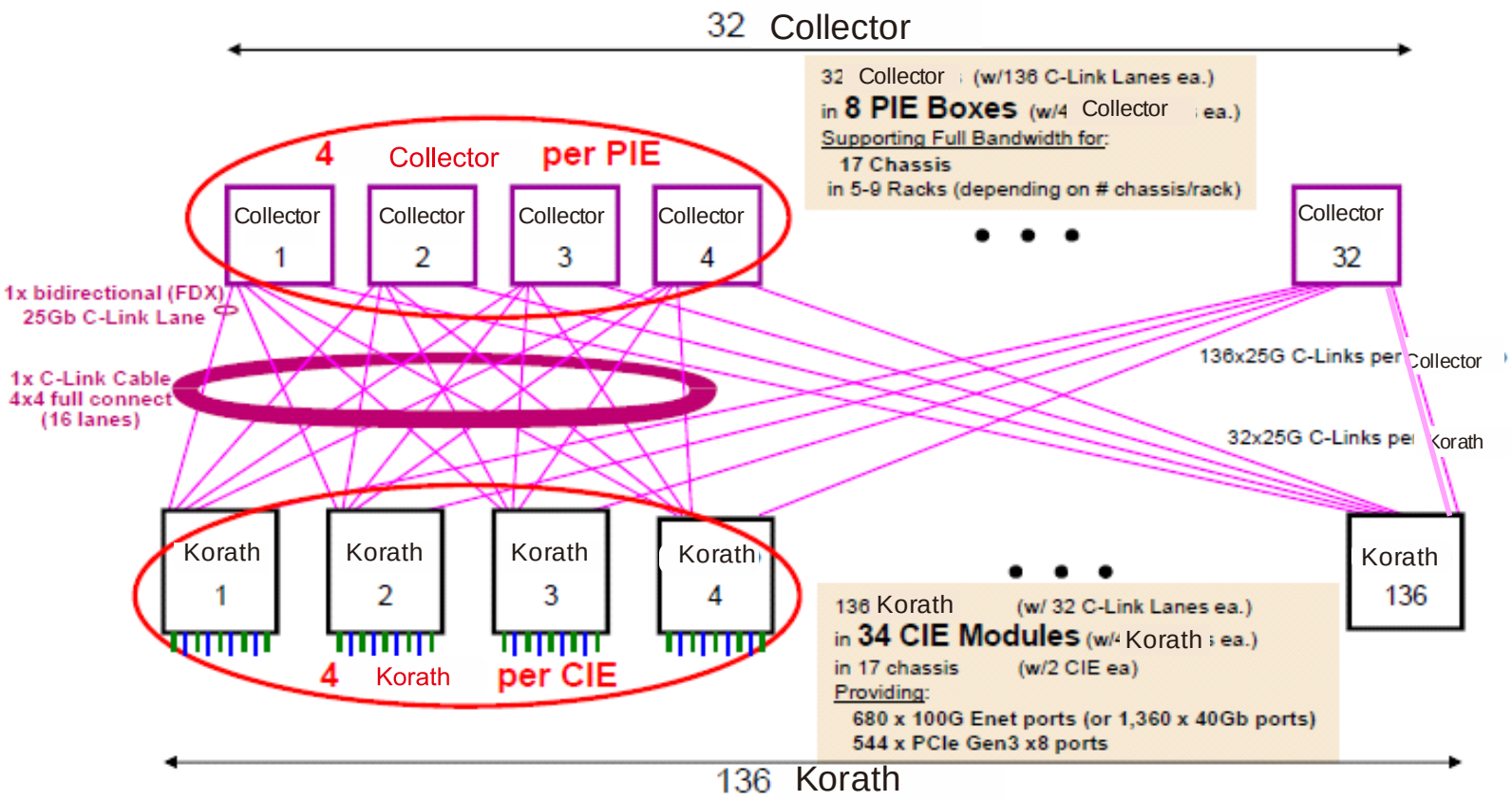


28nm 150W

Nebula应用模式


Ethernet: 5 x 100Gb
PCIe: 4 x Gen3 x8
Korath chip
internal crossbar switch
32 x 25Gb C-links
Collector chip
136 X 136 crossbar switch
136 x 136 (25Gb C-links)
32nm ASICs with 25Gb Serdes
System Configurations
FULL MESH
Best cost/power/density for Smaller configs
SPINE-LEAF
Best scalability for Larger configs
S9 S10 S11 S12 S13 S14 S15 S16 S17
X1 X2 X3 X4 X5 X6 X7 X8
S1 S2 S3 S4 S5 S6 S7 S8

Nebula机柜互联


32 Collector
32 Collector (w/136 C-Link Lanes ea.)
in 8 PIE Boxes (w/4 Collector ea.)
Supporting Full Bandwidth for:
17 Chassis
in 5-9 Racks (depending on # chassis/rack)
4 Collector per PIE
Collector 1
Collector 2
Collector 3
Collector 4
Collector 32
1x bidirectional (FDX) 25Gb C-Link Lane
1x C-Link Cable 4x4 full connect (16 lanes)
136x25G C-Links per Collector
32x25G C-Links per Korath
Korath 1
Korath 2
Korath 3
Korath 4
Korath 136
4 Korath per CIE
136 Korath (w/ 32 C-Link Lanes ea.)
in 34 CIE Modules (w/4 Korath ea.)
in 17 chassis (w/2 CIE ea)
Providing:
680 x 100G Enet ports (or 1,360 x 40Gb ports)
544 x PCIe Gen3 x8 ports
136 Korath

Nebula系统特点

■扁平化的单层分布式交换架构

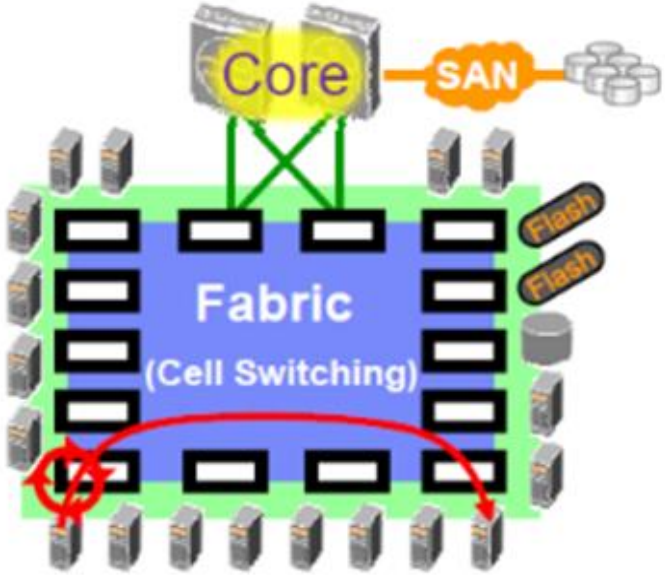
- 高带宽（40Gbps~100Gbps/通道）
- 低延迟（任意两网络节点间一跳可达，且延迟小于1us）
- 性能与服务器、存储节点在网络中的位置无关

■ 更好的支持虚拟机和存储

- 任意位置的虚拟机和存储获得的带宽和延迟一致
- 支持虚拟机灵活迁移和镜像
- 支持超过10万个虚拟机的应用环境
- 可支持RDMA协议用于高效的快速的虚拟机迁移
- 为两个镜像节点提供的高速、低延迟、可保证的通信链路

■ 更好支持SDN

- 大容量多级流表和可编程以太网协议处理引擎
- 将传统的以太网SDN技术引入到PCIE交换中
- 支持目前的主流SDN架构



Nebula数据中心网络

Nebula系统特点（二）

- **芯片级的自适应路由、动态负载均衡以及Qos技术**
  - 硬件自动感知网络状态，避免拥塞，
  - 链路利用率高
  - 网络性能在重负载下保持一致
  - 为存储、通信、计算、IO流量提供可预知的带宽和QoS保证
  - 在设计上，保证了大型的“elephant”流不影响 “mice”流。使得虚拟机迁移、虚拟机镜像、数据库备份、磁盘阵列-闪存阵列交互等应用产生的大数据流不影响延迟敏感的通信应用。
  - 管理便捷，无需专家级人员规划和维护网络负载均衡
- **支持Scale-out可扩展性和灵活性**
  - 每一个机柜就是一个独立系统，多个机柜可以互连成更大规模系统，资源可以跨机柜共享。目前可支持17个机柜，680个100G以太端口和544个PCIE GEM3X8端口，远超传统高端SMP服务器规模。并且还可具备进一步扩大到数百个机柜规模的潜力。
  - 每个机柜内部可以根据需求灵活的配置计算模块、存储模块、PCIE IO模块、闪存模块、网络模块的数量和位置，并且支持动态调整。
  - 达到同等带宽所需的交换机和线缆的数量为传统网络的1/4

Nebula系统特点（三）

■ 丰富的应用协议和高扩展性

- 面向网络应用支持以太网协议，面向存储应用支持FCoE ，面向高性能集群支持ROCE(RDMA over Ethernet)，面向IO和协处理器应用支持PCIE。针对FCoE有专门的优化拥塞避免算法。
- 利用内部的低延迟C-link信元交换技术，将来可方便的引入更多的开放标准接口，如SAS、SCSI、FC等等
- 支持端到端的重传，硬件支持链路上CRC校验与重传

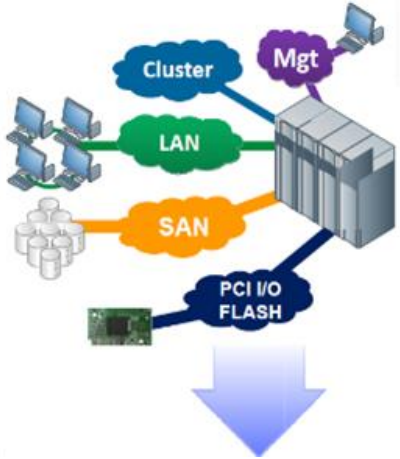
■ 数据中心融合网络

- 针对数据库集群、高性能计算集群，同时满足高带宽的大数据流存储需求和低延迟的消息传递需求
- 一次投入，存储、IO、计算、共享内存、网络等都享有当前最高速的互连带宽，更低的投资、使用和维护成本
- 二次开发成本低，用户可自定义专用协议和接口，一次转换可接入整个系统，例如SAS、SCSI、FC可通过PCIE适配器直接连入系统。

■ 数据中心虚拟化

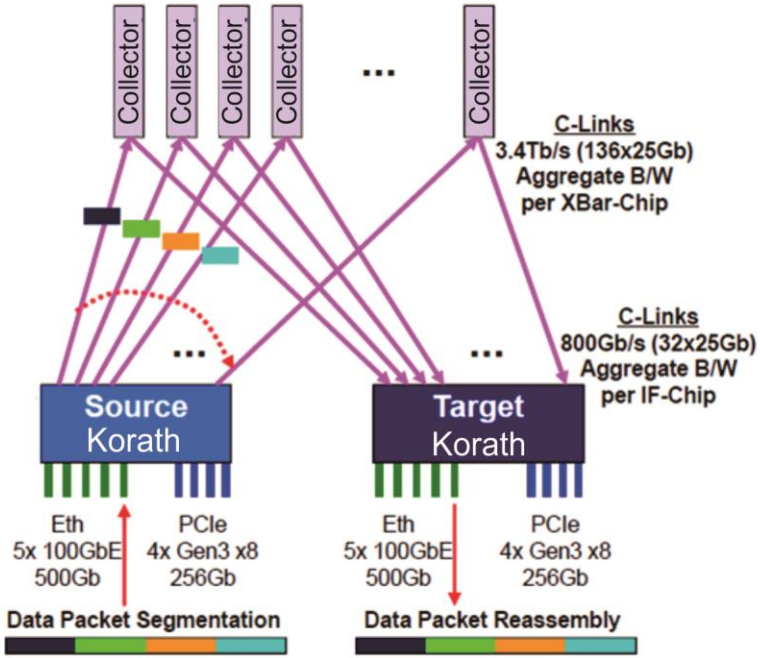
- 支持SR-IOV、MR-IOV，任意节点的PCIE适配器，全系统共享可见，物理共享、虚拟隔离

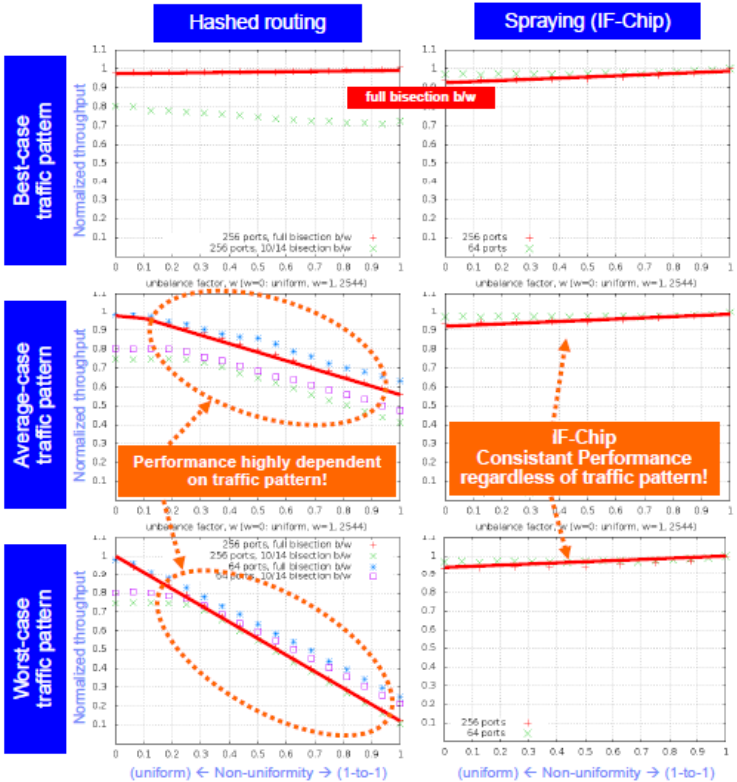
基于传统互联网络的多网络平台



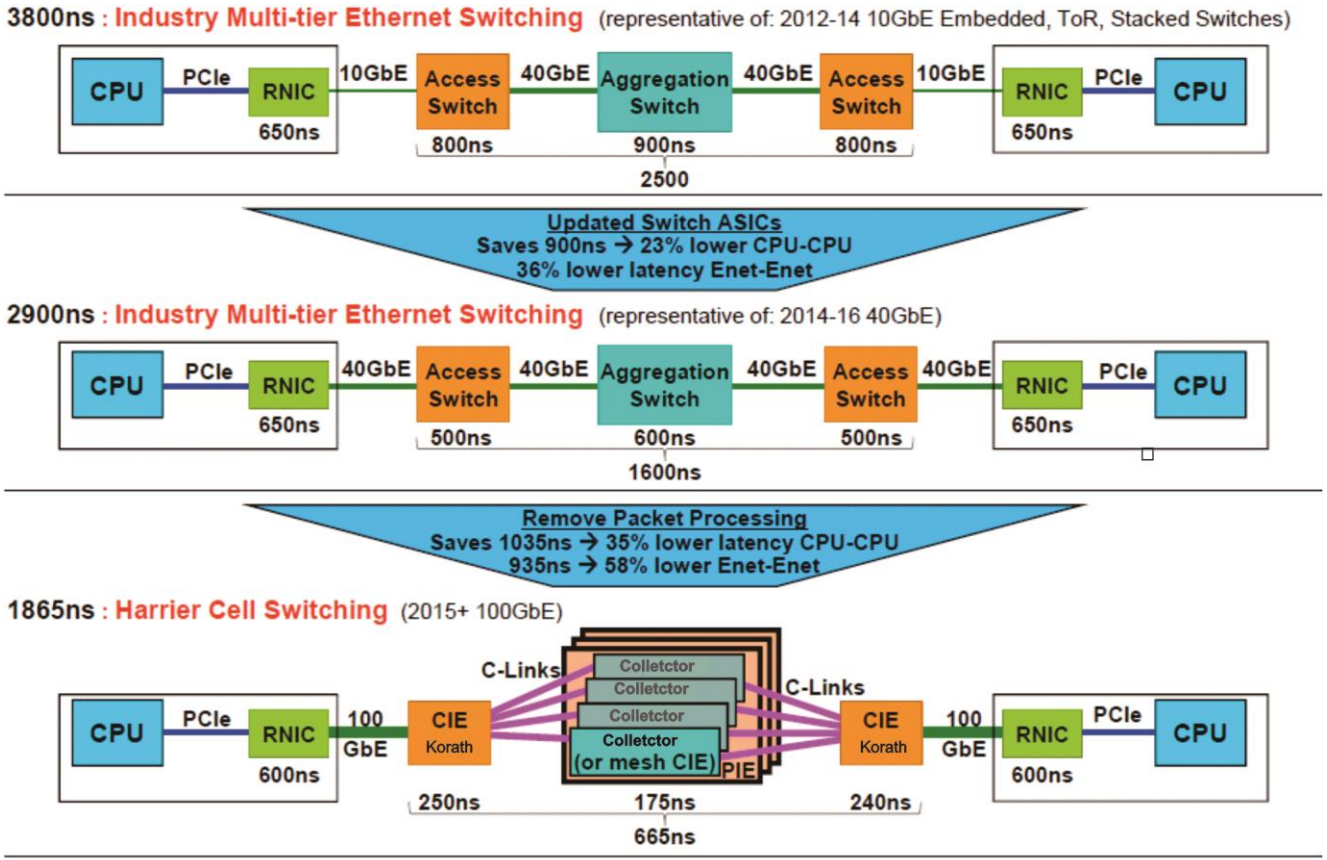
基于Nebula的融合网络平台

Nebula信元交换技术

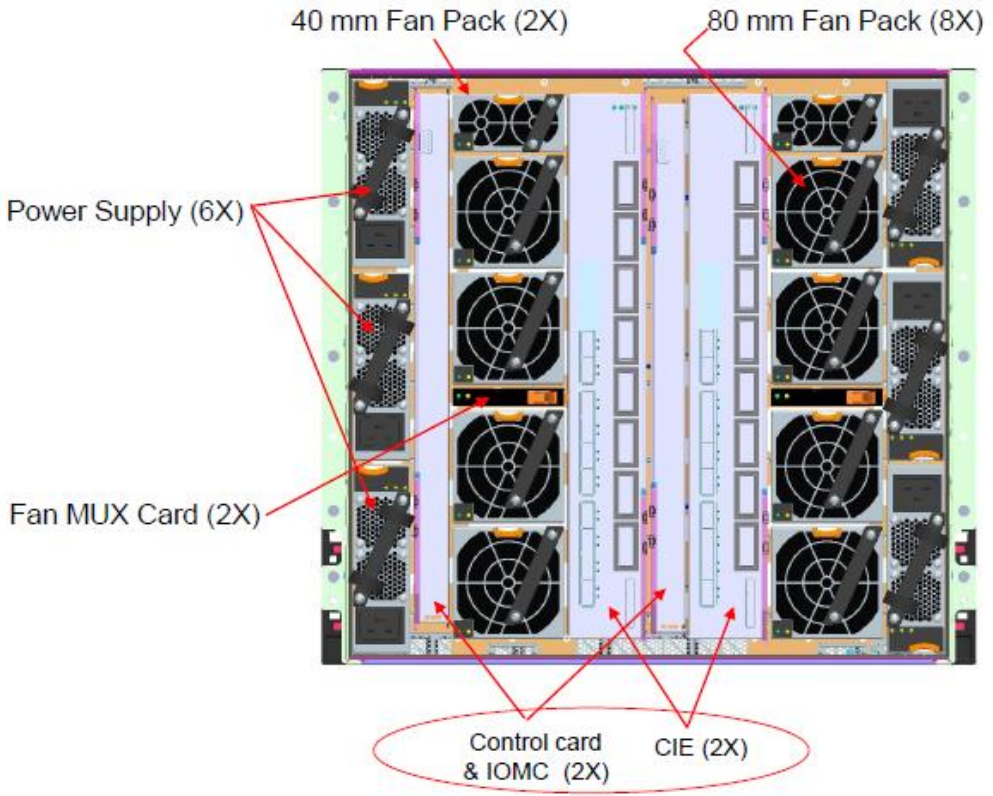




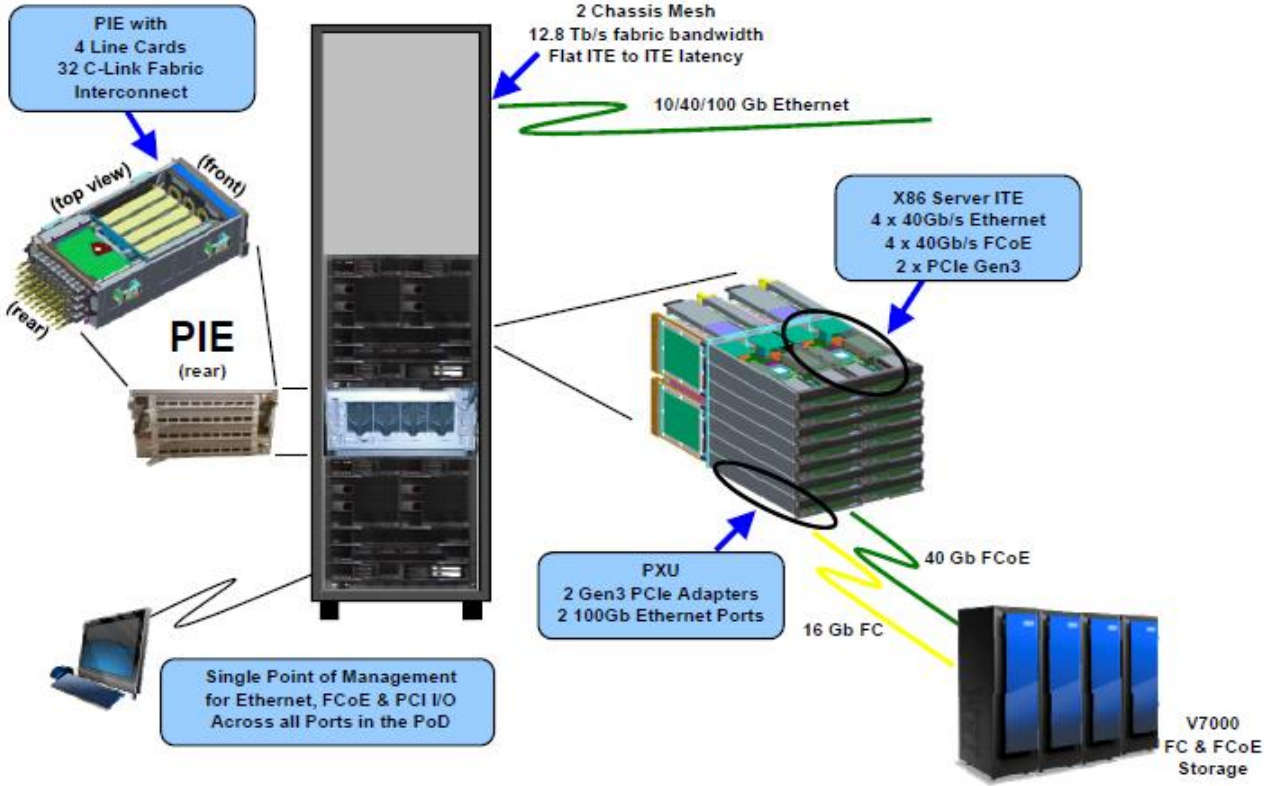
Nebula低延迟转发



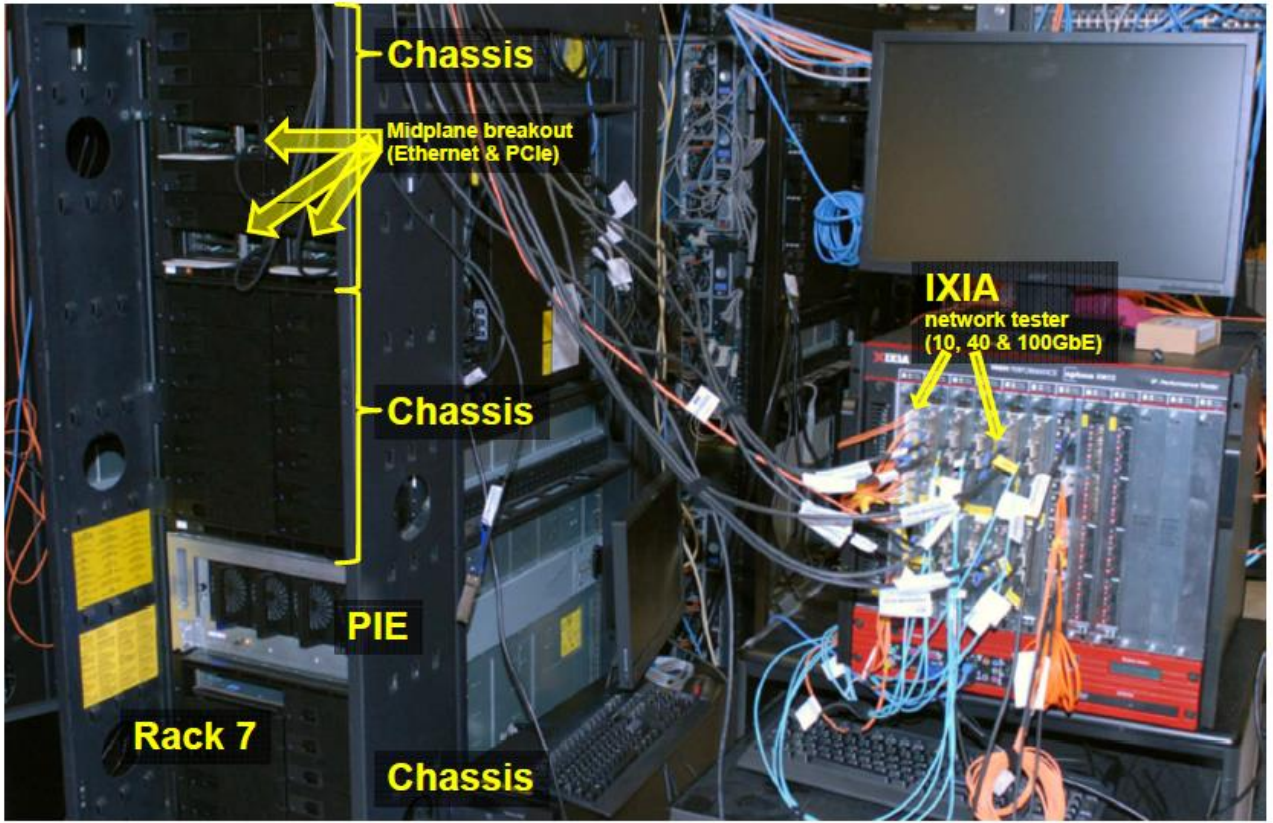
Nebula系统机框


40 mm Fan Pack (2X)
80 mm Fan Pack (8X)
Power Supply (6X)
Fan MUX Card (2X)
Control card
& IOMC (2X)
CIE (2X)

Nebula系统形态


PIE with
4 Line Cards
32 C-Link Fabric
Interconnect
(top view)
(front)
(rear)
PIE
(rear)
2 Chassis Mesh
12.8 Tb/s fabric bandwidth
Flat ITE to ITE latency
10/40/100 Gb Ethernet
X86 Server ITE
4 x 40Gb/s Ethernet
4 x 40Gb/s FCoE
2 x PCIe Gen3
PXU
2 Gen3 PCIe Adapters
2 100Gb Ethernet Ports
40 Gb FCoE
16 Gb FC
V7000
FC & FCoE
Storage
Single Point of Management
for Ethernet, FCoE & PCI I/O
Across all Ports in the PoD

Nebula原型测试系统



Thanks !