Huawei Enterprise A Better Way

云计算时代，无阻塞交换

www.huawei.com

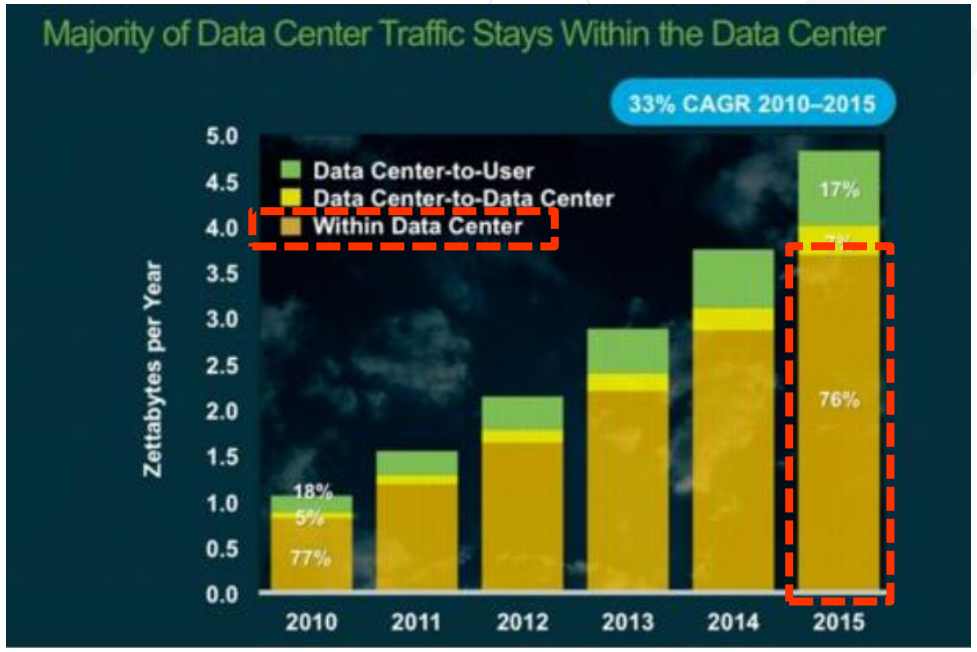
HUAWEI

内容综述

- **通过本胶片，您能够了解到：**

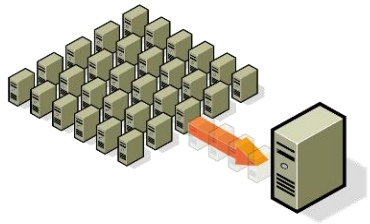
  - **云计算时代，为什么需要无阻塞交换？**
    - 云计算时代，横向流量成为数据中心的主体
  - **我们可以通过怎样的技术来实现无阻塞交换？**
    - 主体技术：胖树架构
    - 辅助技术：高速互联链路、大容量缓存
  - **华为公司面对无阻塞交换的需求，提供了什么样的产品和解决方案？**
    - 全新一代数据中心交换机产品：CE12800、CE6800，全面覆盖数据中心网络需求
    - 全新的数据中心无阻塞交换解决方案，满足云计算需求，提供超大规模服务器集群能力

云计算时代，数据中心的流量模型发生了变化



从预测数据看数据中心的流量

- **资源池化引发资源间横向流量**
  - **VM迁移**：需要把VM内大量的实时数据进行搬移
  - **VM交互**：VM的自由部署，使VM间交互占用网络资源



- **大数据业务带来大量横向流量**
  - 并行计算、3D渲染、搜索、等业务，需要**服务器集群**执行协同运算，导致在数据中心内产生大量交互数据

从业务应用看数据中心的流量

云计算时代，超过70%的流量是数据中心内的东西向流量

大量的东西向流量，驱动了数据中心网络的改变

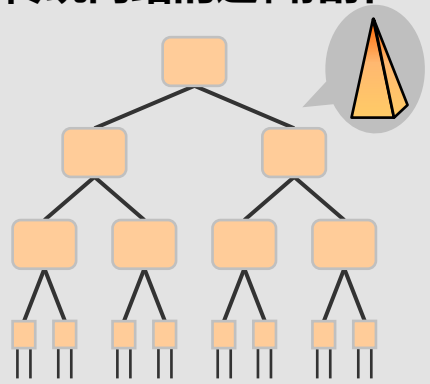
- **传统数据中心，可采用收敛的网络架构**
  - **流量特征**：80%为南北向流量。基于业务特点和出口带宽，流量收敛
  - **网络架构**：**收敛架构**，收敛比为4:1～20:1
- **云计算时代，需采用无阻塞的网络架构**
  - **流量特征**：70%为东西向流量。基于云业务的需求，流量较少收敛
  - **网络架构**：网络不收敛，为**无阻塞的架构**

云计算时代，需要无阻塞的交换网络

实现无阻塞交换的技术手段：胖树架构

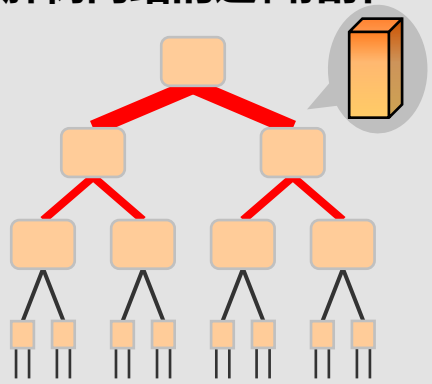
胖树网络和传统网络的比较

传统网络的逻辑拓扑



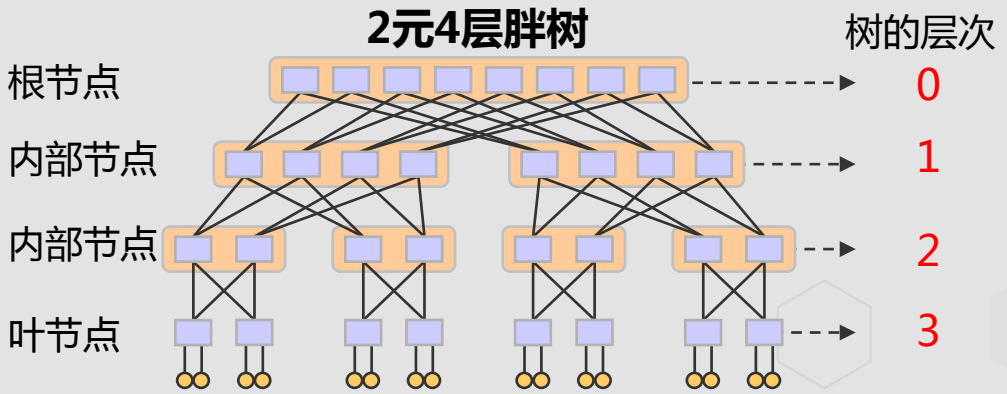
传统树形网络，树根处和叶子处的枝干粗细相同（链路带宽相同），即**从叶子到树根，链路带宽逐层收敛**

胖树网络的逻辑拓扑



胖树网络拓扑，更像真实的树，越到树根、枝干越粗（链路带宽越大），即**从叶子到树根，链路带宽不收敛**

胖树技术所描述的物理拓扑

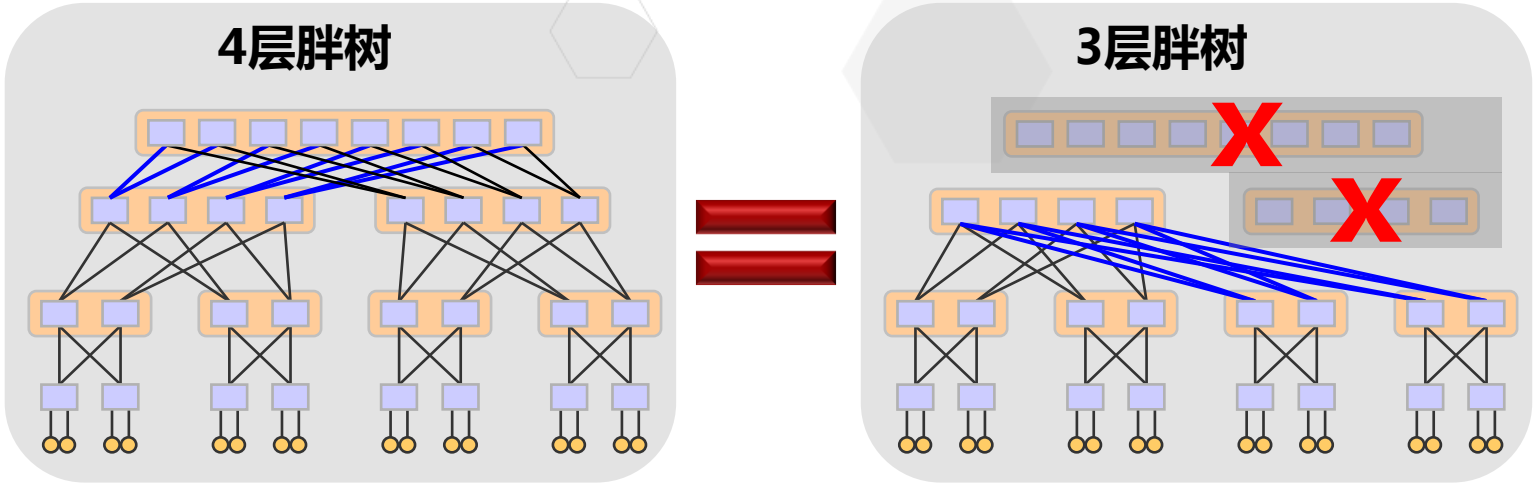


在每一层的每台网络设备都相同的前提下：

- **交换单元**：胖树每一层所需的交换单元数量相同(根节点不需要向上互联，则可以除外)，即**每层具有相同的交换能力**
- **物理连线**：胖树每一层所需的物理连线数量相同，即**每层的出线带宽相同**

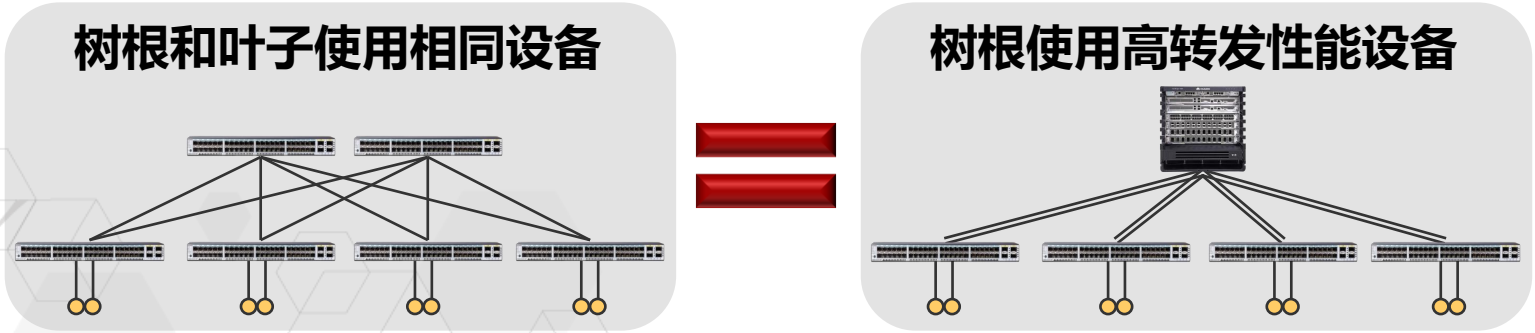
胖树架构的核心在于：胖树内每层网络的上下行带宽保持一致

胖树架构在实际使用中的演变



树根不需要向上扩展

- **实施**：把树根计划用于上连的端口也接入到胖树中
- **效果**：胖树减少一层，树根交换机减少一半，网络布线简化

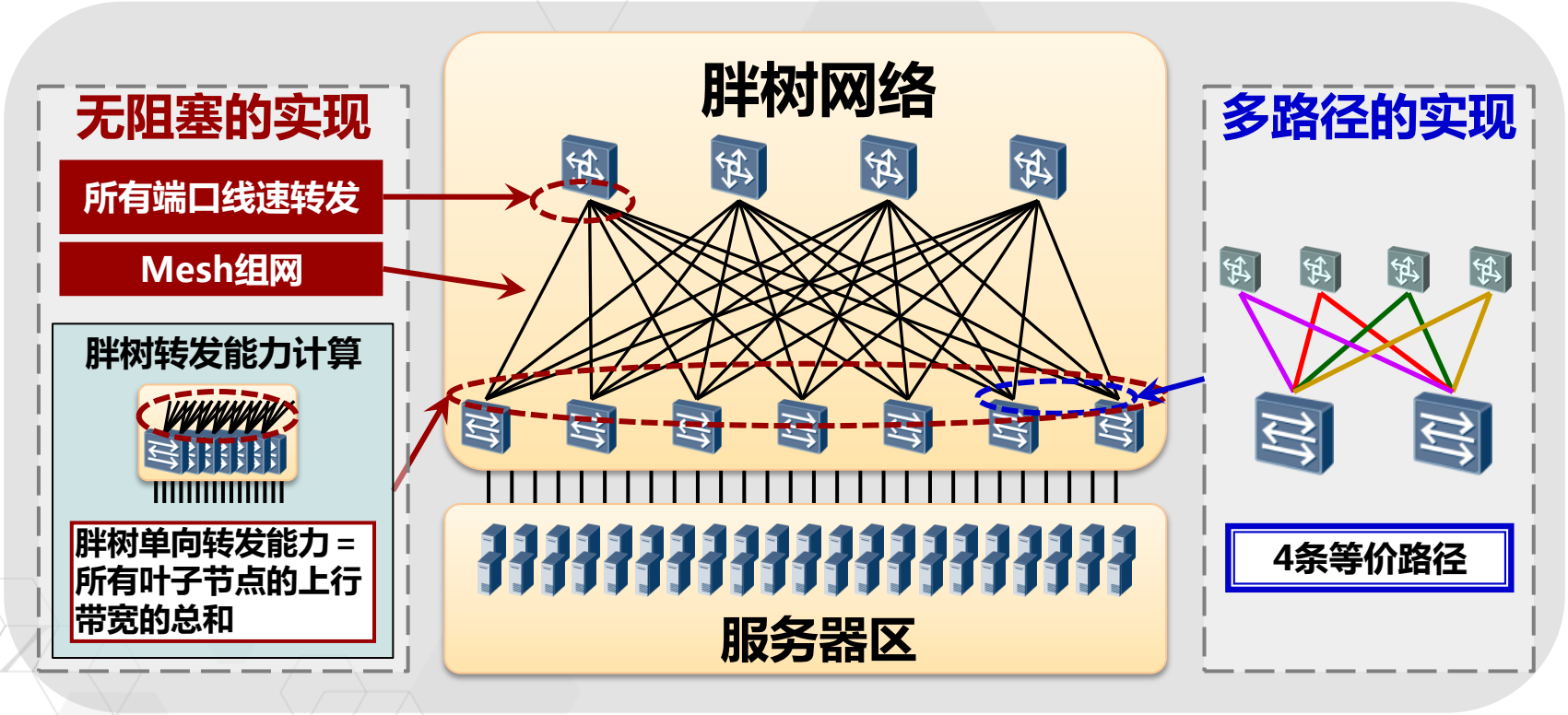


树根使用高性能设备

- **实施**：树根被换成高性能设备（如：框式交换机）
- **效果**：减少设备数量，降低网络复杂度

通过优化，带来成本的降低、性能的提升、部署的简化

胖树架构在数据中心的部署方法



无阻塞的部署：胖树物理拓扑

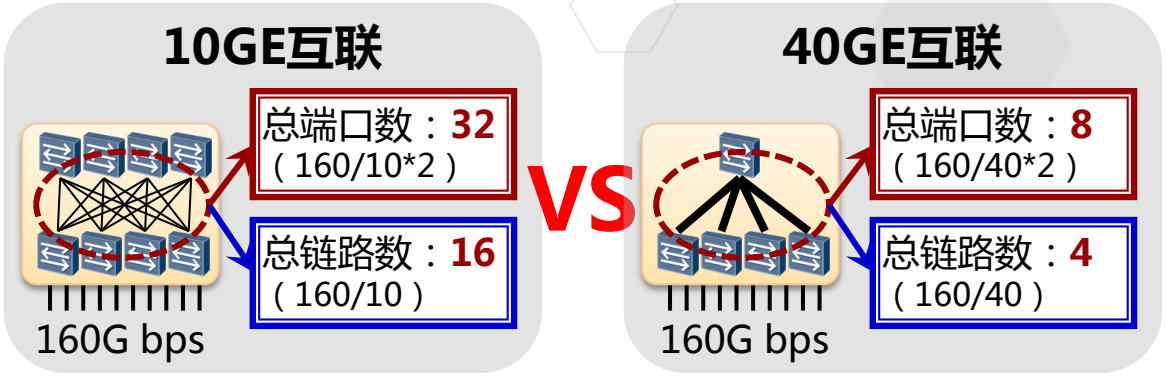
- **网络层次**：二层胖树，或多层胖树
- **网络布线**：叶子和上层设备间Mesh互联，其他层保持出线带宽和叶子相同

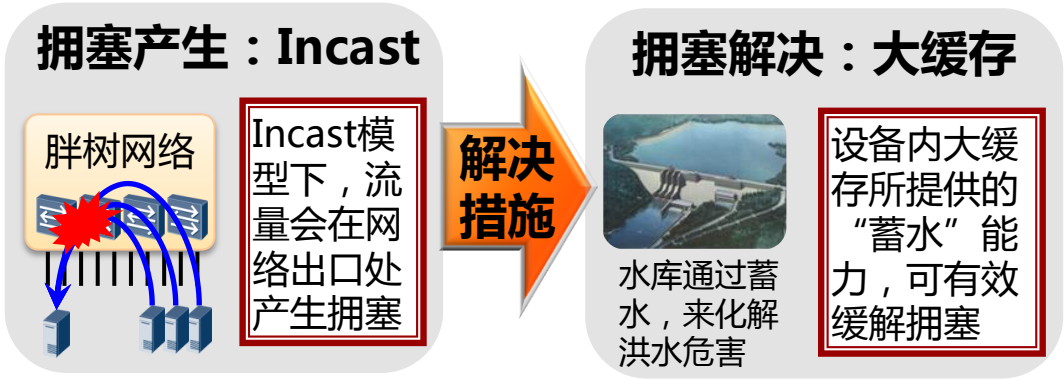
无阻塞的使用：多路径协议

- **环路问题**：胖树多路径的充分使用，会产生环路
- **问题解决**：通过多路径协议来执行ECMP。如：TRILL、OSPF、ISIS

胖树拓扑+多路径协议，构成完备的无阻塞网络

胖树架构的左膀右臂：高速互联链路+大容量缓存





高速互联链路：更强大的武器，让网络高效运转

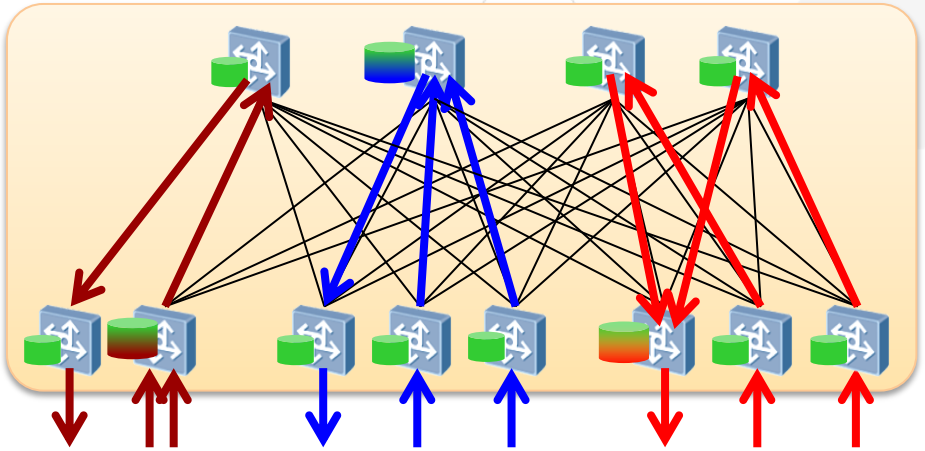
- 上图是提供160G转发能力的二层胖树的例子。可以看出，使用高速互联链路带来巨大的收益：
  - **节约端口**：减少使用的端口数，进而减少所需的设备数，降低成本和部署复杂度
  - **简化布线**：使用较少的物理链路实现相同的性能，降低布线成本和布线复杂度

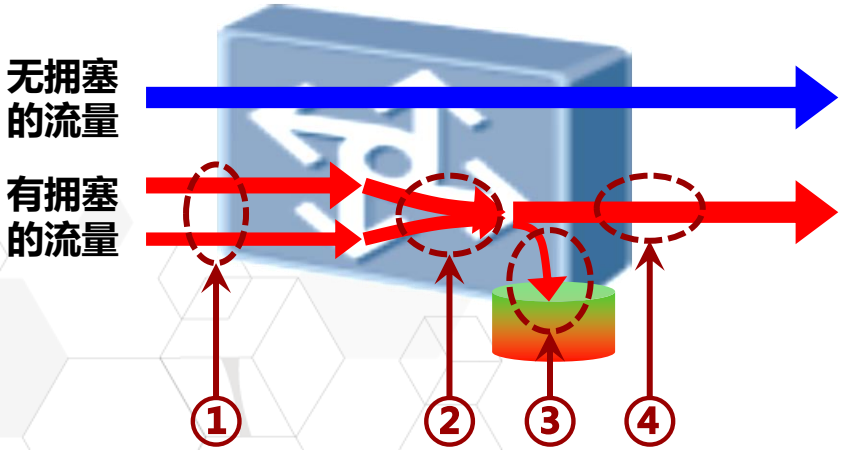
大容量缓存：更坚固的防守，助网络抵挡拥塞

- **什么是Incast流量模型**：**多打一**的流量模型，即多台服务器同时向同一台服务器发送流量。Incast会导致拥塞，并且无法通过网络设计来避免
- **Incast流量拥塞的解决**：设备内集成大容量缓存，在发生拥塞时对流量进行缓存，待链路空闲时再发送出去，可以较好的解决此类拥塞问题

攻守兼备的胖树网络的两大保证：高速互联链路+大容量缓存

缓存的部署和使用





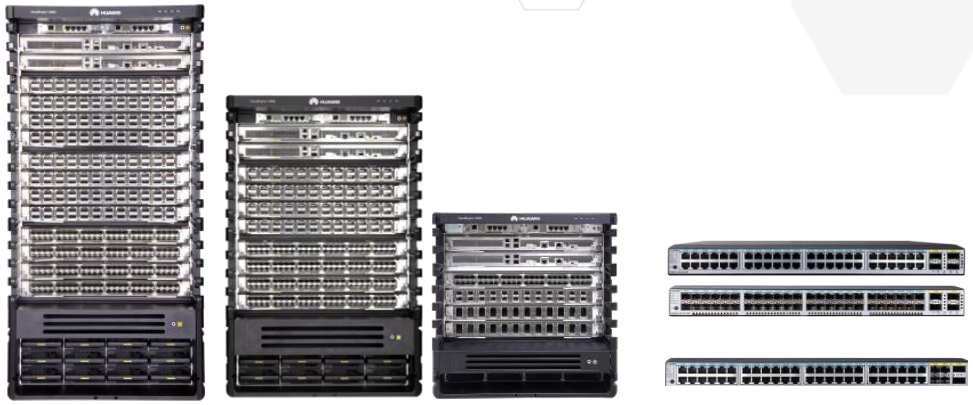
缓存的部署：

- **识别拥塞**：分析网络拓扑和业务模型，识别出网络内可能产生拥塞的位置
- **部署缓存**：在各个可能产生拥塞的位置，都部署具备缓存能力的交换机

缓存的使用：

- **无拥塞的流量**：在交换机内线速转发。缓存的存在不会影响报文的转发
- **有拥塞的流量**：出端口处拥塞，处理过程是：
  - ① 多个端口向一个端口发送报文
  - ② 流量在出端口处汇聚，产生拥塞
  - ③ 基于QoS的调度，把超过端口转发能力的报文缓存起来，待出端口空闲时再转发出去
  - ④ 能够正常转发的报文，被出端口全速转发

华为面向云计算时代的交换机产品和无阻塞解决方案



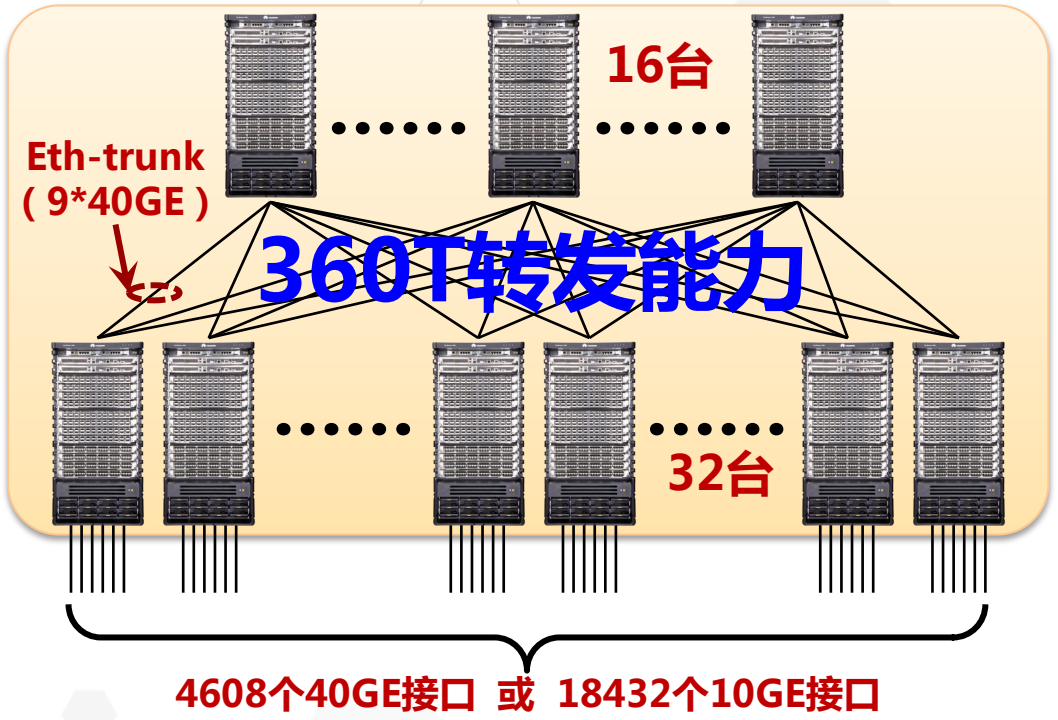
新一代的高性能数据中心交换机产品

- 面向云计算，覆盖高端/低端、框式/盒式各种需求：
  - **CE12800系列框式交换机**：业界最佳高密度板卡设计，支持48*10GE、24*40GE全线速板卡
  - **CE6800系列盒式交换机**：满足各种应用需求，提供GE/10GE下行、10GE/40GE上行

灵活的无阻塞交换的解决方案

- CE12800和CE6800全面支持无阻塞网络部署：
  - **框式交换机组成无阻塞网络**：当前可提供高达360T的双向转发能力，未来可支持更高性能
  - **框式和盒式交换机组成无阻塞网络**：适应TOR组网方案，把无阻塞网络延伸到服务器机架

框式交换机构建无阻塞网络



无阻塞转发能力：
9*40G（bps/叶子和根）*32（叶子）*16（根）*2（双向）= **360T** bps
接入能力：
40GE：24（个40GE/叶子槽位）*6（槽位）*32（叶子） = **4608**个40GE
10GE：4608（个40GE）*4（把40GE拆分成4个10GE） = **18432**个10GE

高性能无阻塞胖树架构

- 大规模胖树组网，配合TRILL协议的多路径能力，组成无阻塞交换网络
- 配合高性能CE12800交换机，网络内可提供高达**360T**的无阻塞双向转发能力

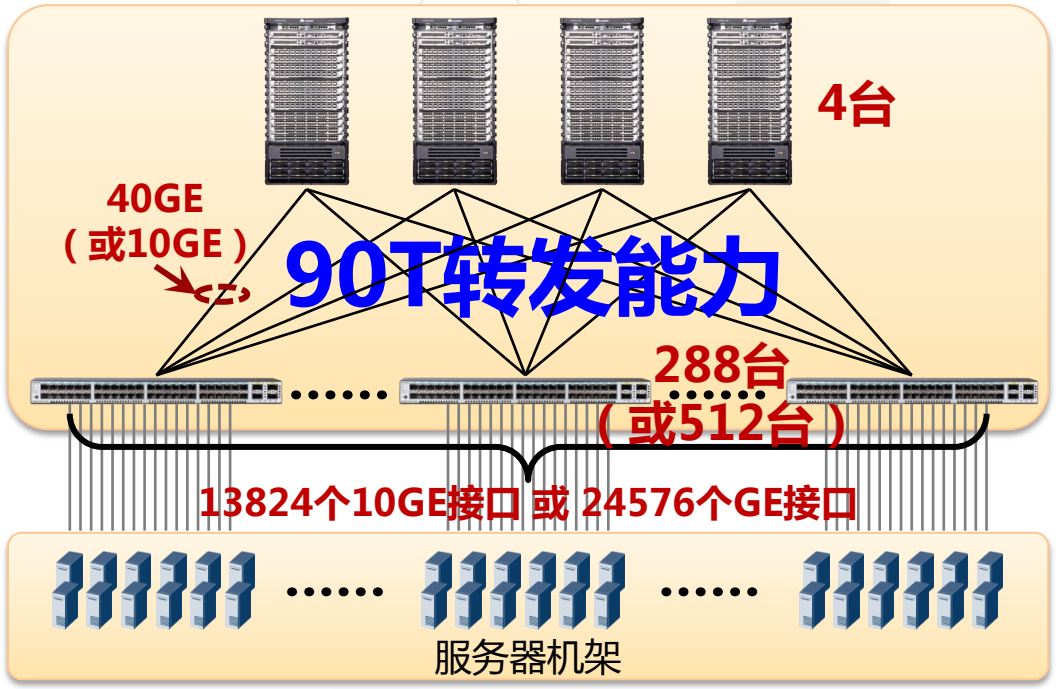
全线速40GE互联

- CE12800提供高密度的40GE线卡：**24*40GE**全线速转发线卡
- 胖树网络内使用40GE线速端口互联，提供高性能无阻塞转发能力

大规模服务器集群的接入能力

- 通过32台胖树叶子交换机，可对外提供高达**4608个40GE**全线速端口
- 通过把40GE端口拆分成4个10GE，可对外提供高达**18432个10GE**全线速端口

框式和盒式交换机构建无阻塞网络



无阻塞转发能力：
40G（bps/叶子和根）*288（叶子）*4（根）*2（双向）= **90T** bps
接入能力：
10GE：48（个10GE/叶子）*288（叶子） = **13824**个10GE
1GE：48（个GE/叶子）*512（叶子） = **24576**个GE
（注：对于GE服务器，S12800使用96*10GE的线卡和TOR互联，可部署512台TOR）

延伸到服务器机架的无阻塞胖树架构

- 通过CE6800，使**胖树架构延伸至TOR**，配合TRILL实现扁平化的无阻塞交换网络
- CE12800和CE6800混合组网，胖树网络内可提供高达**90T**的双向转发能力

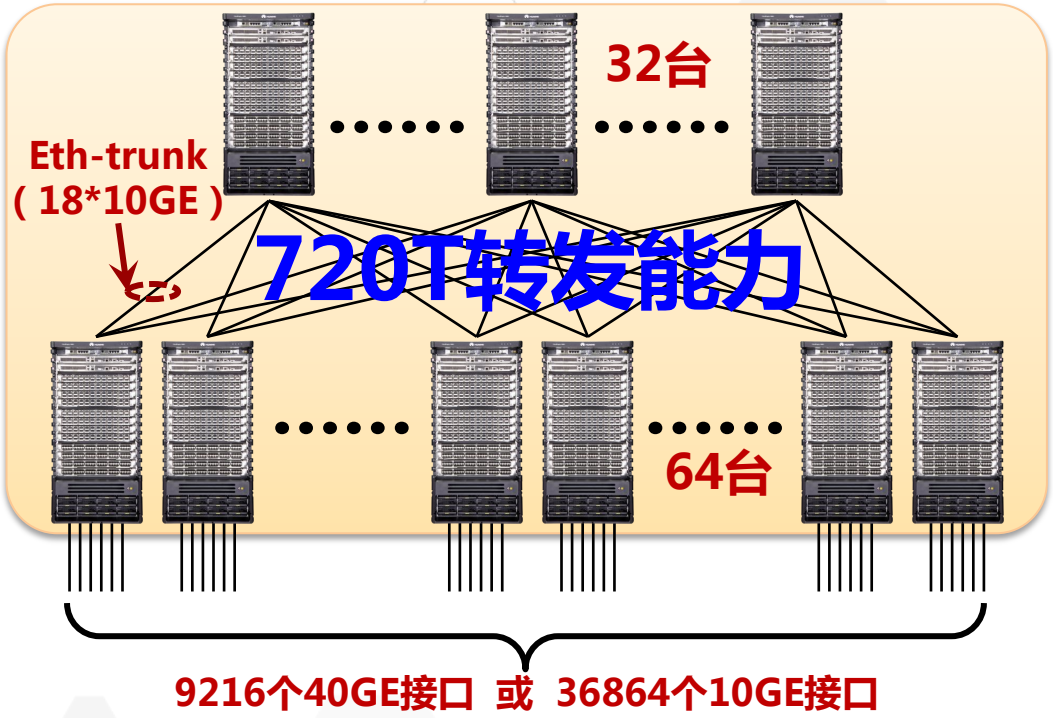
全线速40GE互联

- CE12800提供**24*40GE**的线速高密线卡用于和下行的CE6800对接
- CE6800提供**4*40GE**的线速上行接口和胖树树根的CE12800对接

大规模服务器集群的接入能力

- CE6800下行可提供48*10GE的接入端口，整网可接入**13824台10GE服务器**
- 对于接入GE服务器的场景，网络内交换机之间使用10GE端口互连，基于整网性能的考虑，整网部署512台TOR，可接入**24576台GE服务器**

扩展组网1：框式交换机间10GE互联



无阻塞转发能力：
18*10G（bps/叶子和根）*64（叶子）*32（根）*2（双向）= **720T** bps
接入能力：
40GE：24（个40GE/叶子槽位）*6（槽位）*64（叶子） = **9216**个40GE
10GE：9216（个40GE）*4（把40GE拆分成4个10GE） = **36864**个10GE

高性能无阻塞胖树架构

- 大规模胖树组网，配合TRILL多达**32路ECMP**的能力，组成更大规模无阻塞网络
- 配合高性能CE12800交换机，网络内可提供高达**720T**的无阻塞双向转发能力

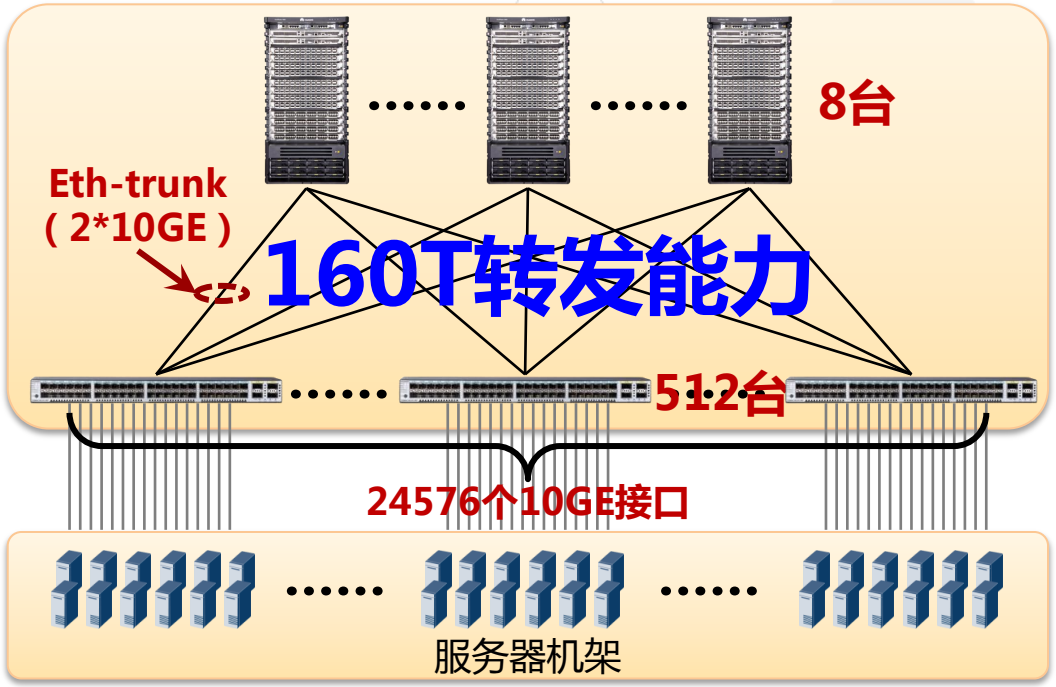
高密度10GE线卡互联

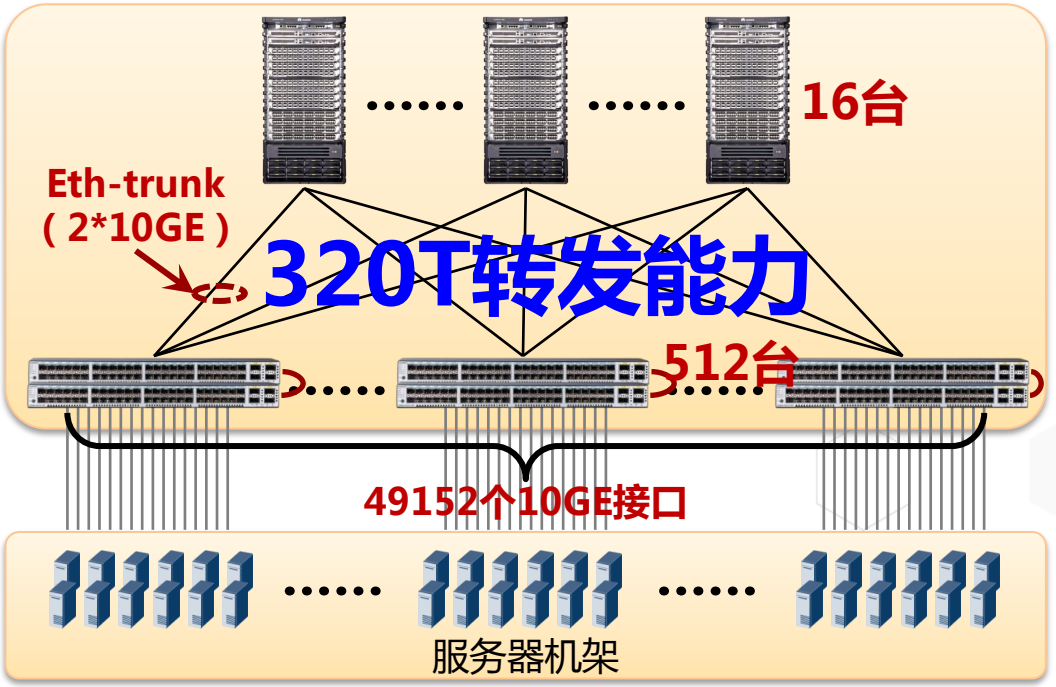
- CE12800提供的**24*40GE**的线速高密线卡，可扩展成**96*10GE**端口
- 配合CE12800的多达**32路链路聚合**，提供胖树网络内高密度的10GE互联

大规模服务器集群的接入能力

- 通过64台胖树叶子交换机，可对外提供高达**9216个40GE**全线速端口
- 通过把40GE端口拆分成4个10GE，可对外提供高达**36864个10GE**全线速端口

扩展组网2：框式和盒式交换机间10GE互联





无阻塞转发能力：
2*10G（bps/叶子和根）*512（叶子）*8（根）*2（双向）= **160T** bps
10GE接入能力：
48（个10GE/叶子）*512（叶子）= **24576**个10GE

无阻塞转发能力：
2*10G（bps/叶子和根）*512（叶子）*16（根）*2（双向）= **320T** bps
10GE接入能力：
2*48（个10GE/叶子）*512（叶子）= **49152**个10GE

Huawei Enterprise A Better Way