



## 超大流量的基础架构保障核心

宋庆春

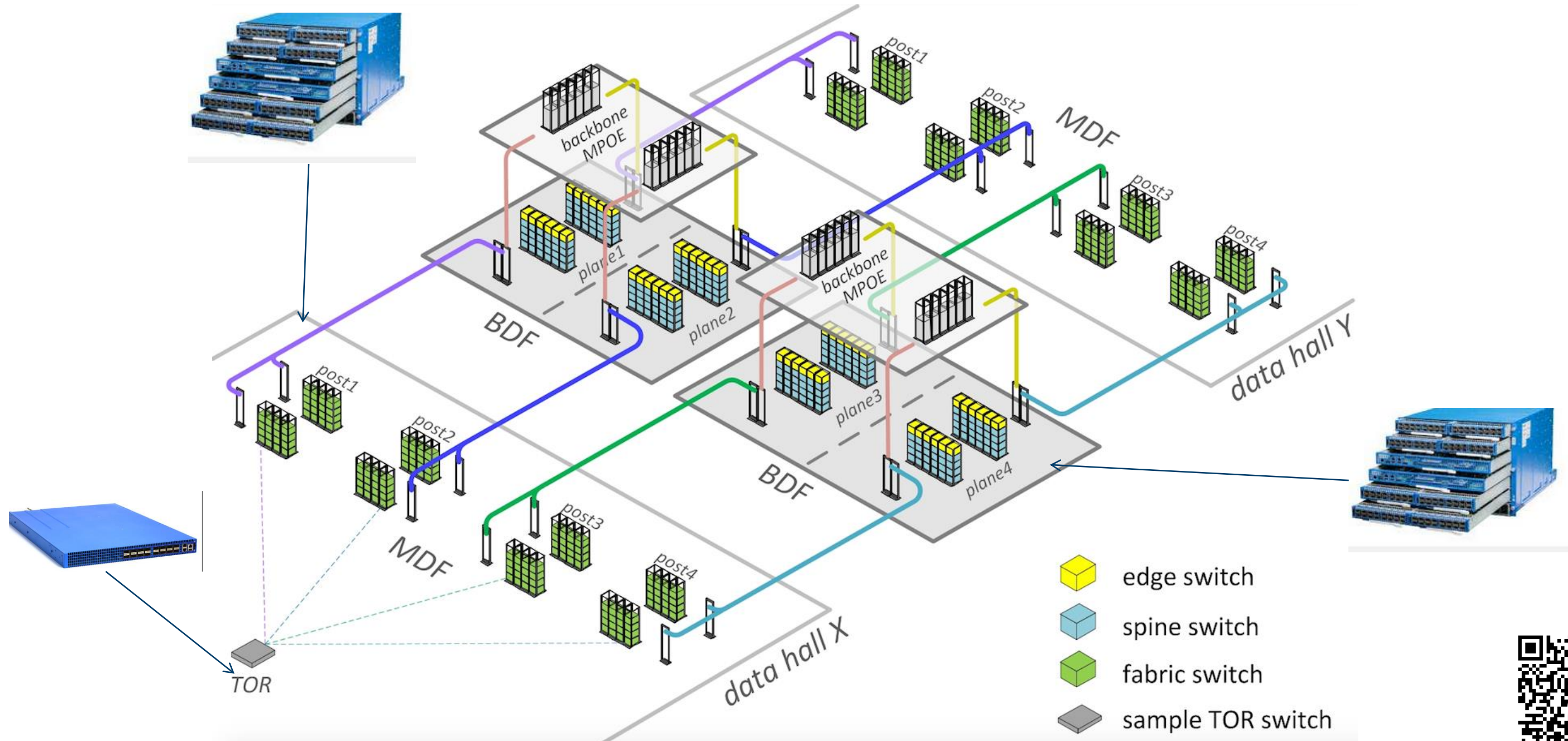
 **Mellanox**  
TECHNOLOGIES  
Connect. Accelerate. Outperform.



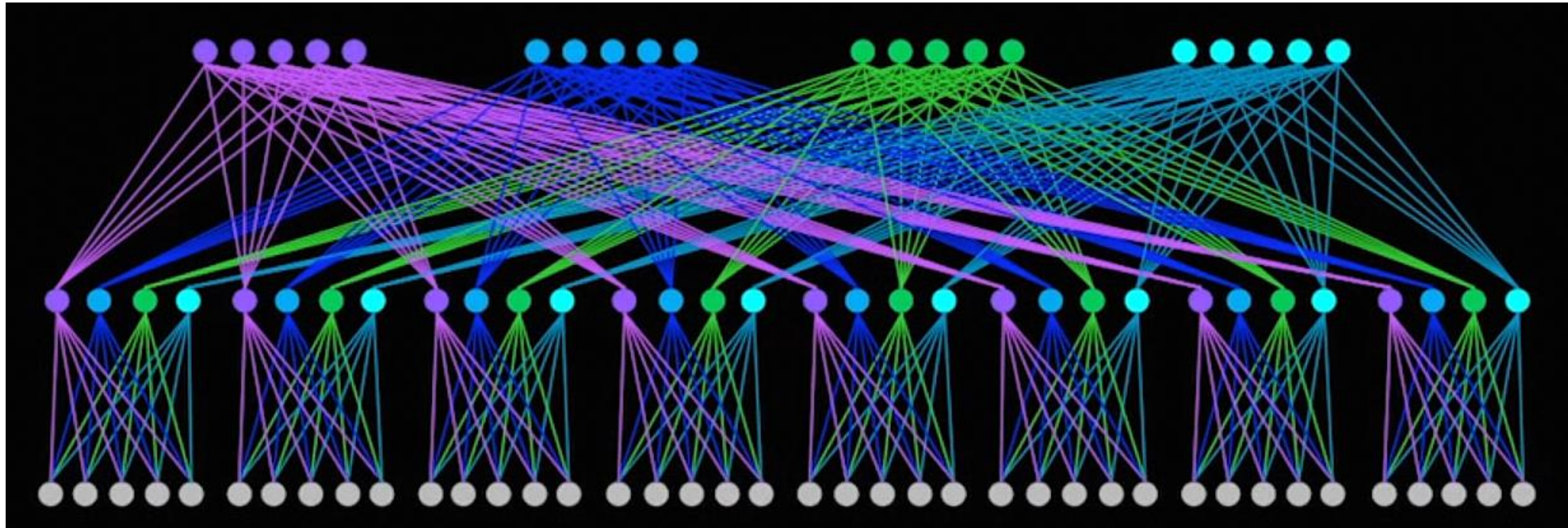
微信扫码收听演讲音频



# Facebook 公司超大规模数据中心一览







## 挑战：

- 如何快速部署超大规模数据中心？单一数据中心超过5万台服务器，百万台虚拟机...
- 交换机，服务器，线缆，虚拟机...
- 如何适应数据中心的渐进增长？从1万台到2万台，从2万台到5万台...
- 如何保证应用的性能？RDMA应用，TCP应用...
- ○ ○ ○ ○ ○ ○



微信扫码收听演讲音频

# 如何快速部署超大规模数据中心？

## ■ 设备模块化

- 使用业界标准服务器，交换机，线缆

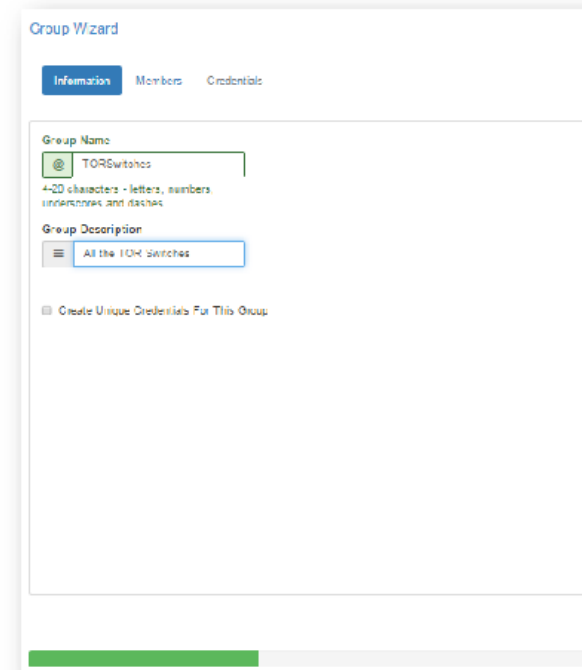
## ■ 软件模块化

- 交换软件标准化，模块化，分层化
  - ToR交换
  - Fabric交换
  - Spine交换
  - Edge交换
- 服务器软件标准化，模块化
  - PXE/UEFI
  - 操作系统
  - OVS/Hypervisor

## ■ ZTP(Zero Touch Provision)

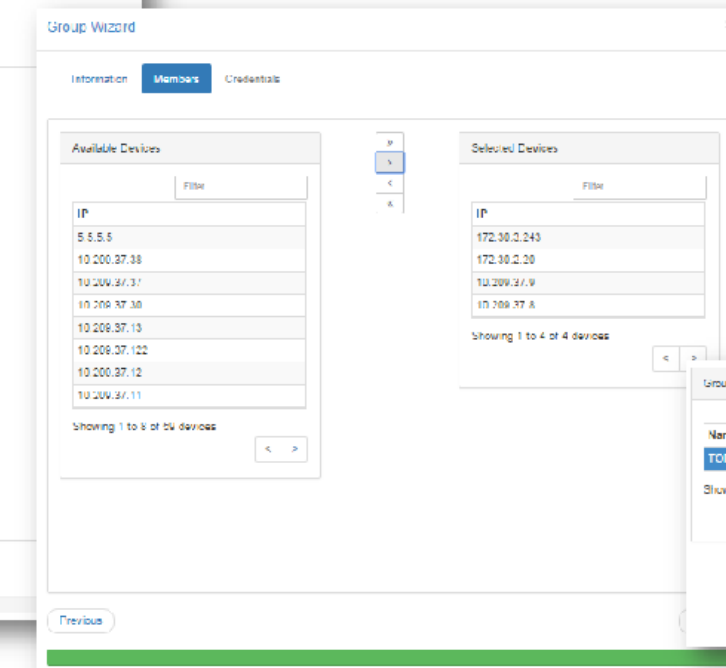
- 软件模块模板化
- 系统部署自动化，减少人为操作

### 1. Create a new group



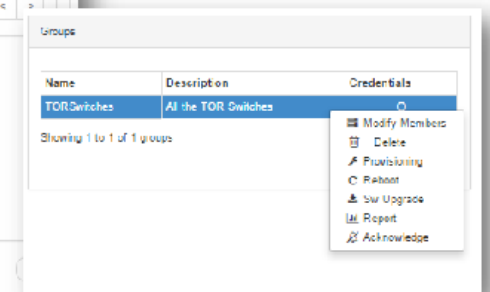
Group Name: TORSwitches  
4-20 characters - letters, numbers, underscores and dashes  
Group Description: All the TOR Switches  
Create Unique Credentials For This Group

### 2. Assign Members



Available Devices: 5.5.5.5, 10.200.37.38, 10.200.37.37, 10.200.37.30, 10.200.37.13, 10.200.37.122, 10.200.37.12, 10.200.37.11  
Selected Devices: 172.30.0.243, 172.30.0.20, 10.200.37.9, 10.200.37.8  
Showing 1 to 4 of 4 devices

### 3. Quick Action Select



Name	Description	Credentials
TORSwitches	All the TOR Switches	0

Showing 1 to 1 of 1 groups

- Modify Members
- Delete
- Provisioning
- Reboot
- SW Upgrade
- Report
- Acknowledge



OPEN  
Compute Project



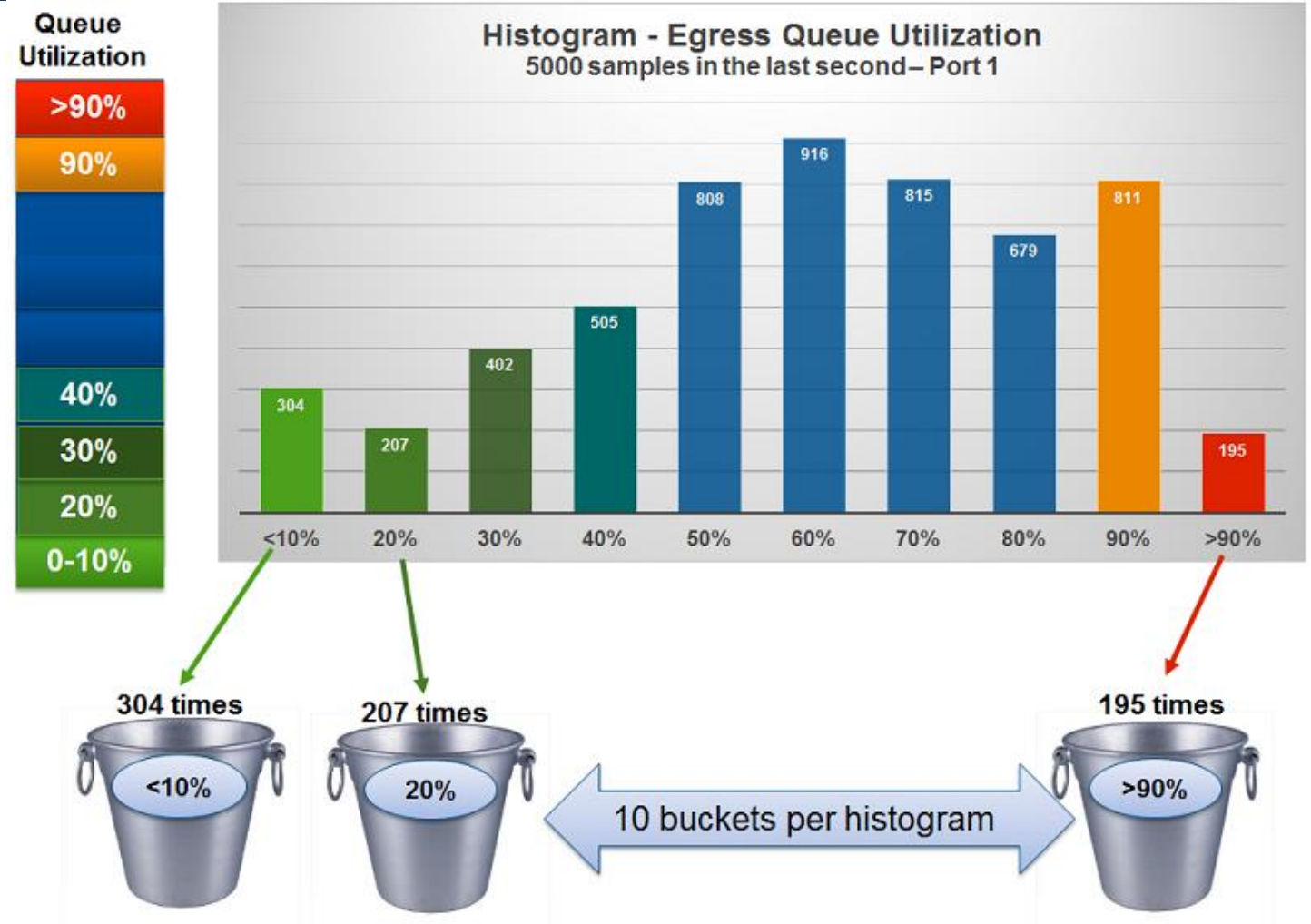
开放数据中心委员会  
Open Data Center Commi





# 如何管理和监控超大规模数据中心？

- 设备 & 软件标准化、模块化
  - 使用业界标准服务器，交换机，线缆
  - 标准软件管理接口，Restful，JSON
- 管理软件一体化
  - 单一管理软件
    - 交换机、网络
    - 服务器、Hypervisor
    - 线缆、光模块
    - 虚机
  - 例如服务器路由化（Routing-In-Host）
- 优化Telemetry
  - 面向Hop
  - 面向TCP Flow
  - 面向RDMA QP
  - 创建历史记录
- 支持任务预先布置(Task Scheduling)
  - 预定义任务执行时间

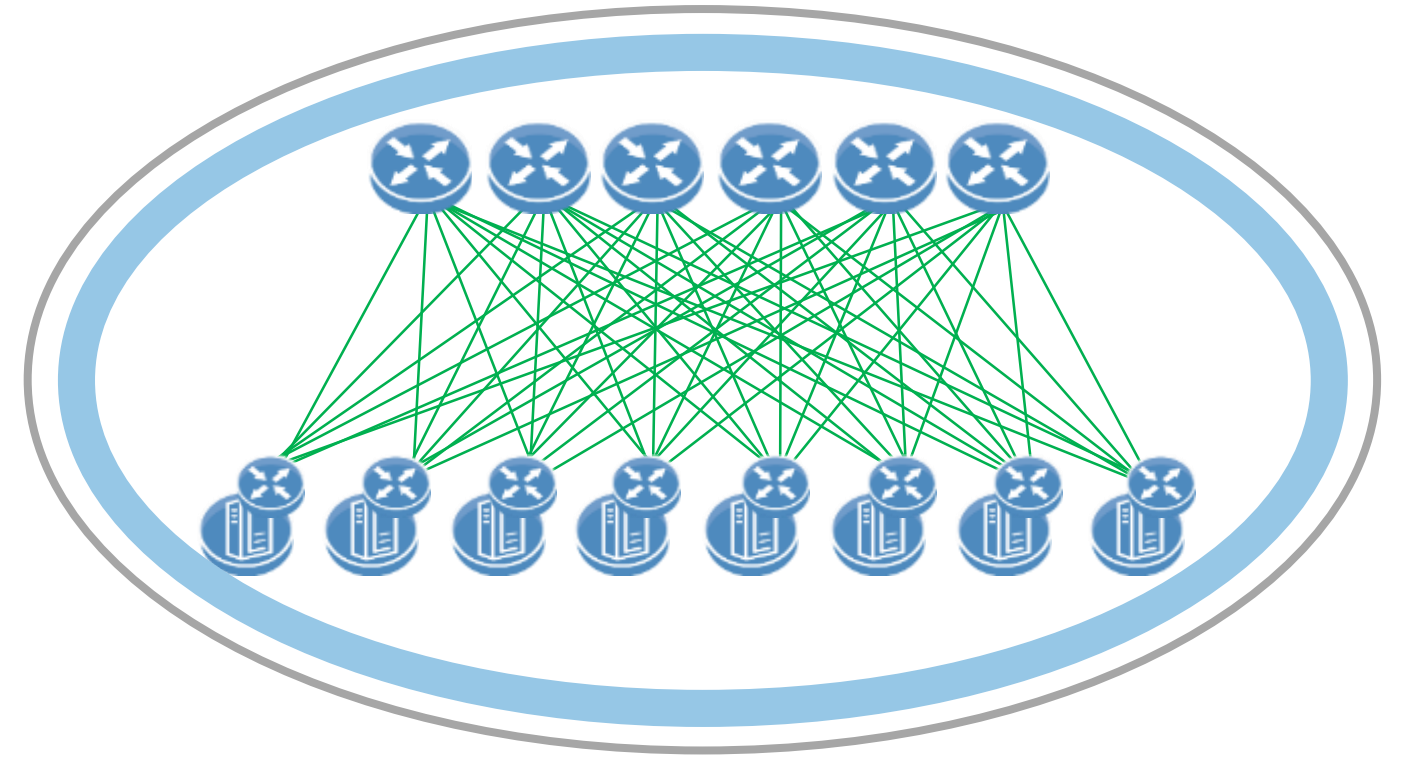


开放数据中心委员会  
Open Data Center Consortium



# 如何适应数据中心的渐进增长？

- 设备 & 软件标准化、模块化
  - 使用业界标准服务器，交换机，线缆
  - 避免新旧设备异构
- 使用CLOS架构
  - Spine + Leaf 架构
    - ToR <-> Fabric switch <-> Spine switch
    - ToR <-> Leaf <-> Spine <-> Super Spine
- 使用标准的网络协议
  - ECMP、OSPF、BGP...
  - EVPN(Controllerless )
    - 避免对SDN控制器的依赖性
- 使用基于标准协议的机框式交换机
  - Facebook 128端口交换机
  - 百度 128端口交换机



**开放数据中心委员会**  
Open Data Center Committee



# 如何保证应用的性能？

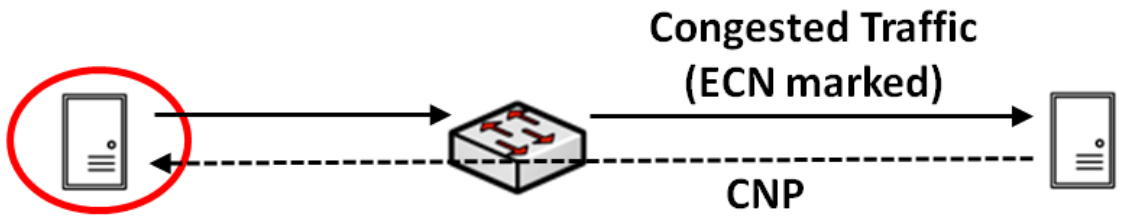
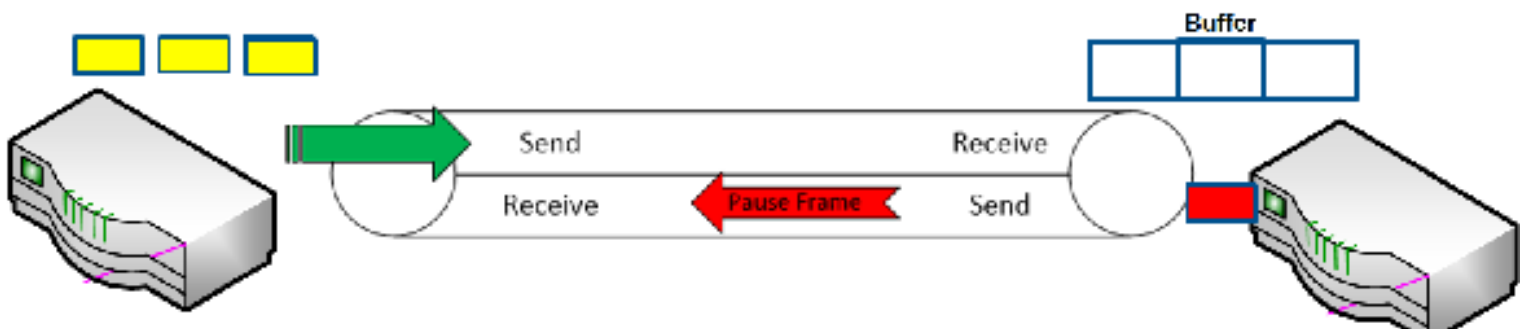


## ■ 应用的发展趋势：

- RDMA + TCP
- 物理机到虚机
- 虚机到物理机（ Bare Mental ）

## ■ 应对措施：

- 设备简单化，够用就好。
  - 无用的软件和硬件会增加出问题的机会
  - 简化硬件服务器、交换机（白牌机）
  - 简化操作系统，通讯方式（RDMA）
  - 有效利用CPU、内存和网络资源
- RDMA网络和TCP网络的隔离
  - 避免RDMA流量TCP流量的互相影响
  - 不同的应用配置不同的优先级
  - 根据使用场景选用不同的网络无损、有损模式
- 有效的虚机迁移机制



**Sender NIC**  
Reaction Point (RP)

**Switch**  
Congestion Point

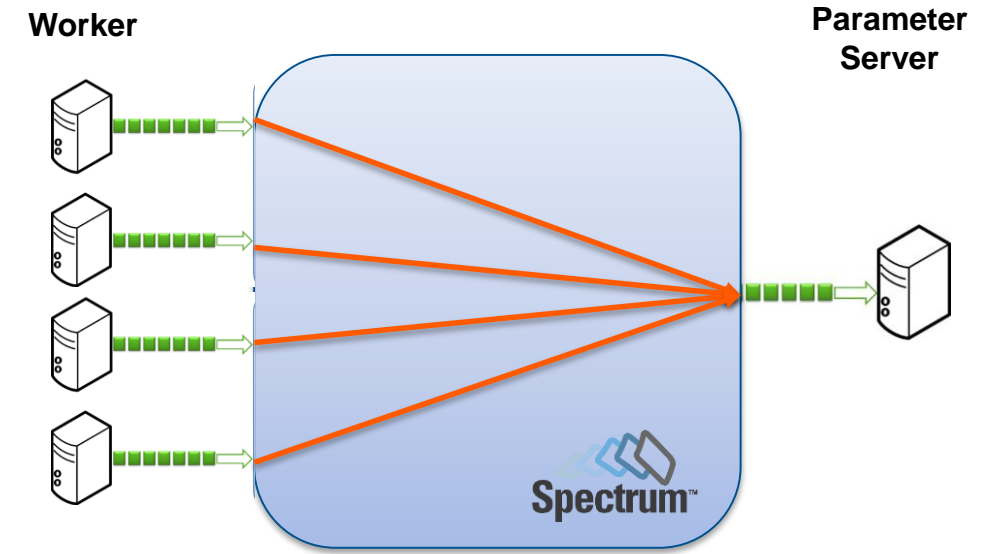
**Receiver NIC**  
Notification Point (NP)

	Lossy RoCE	Lossy w/ QoS	Semi-Lossless	Lossless
ECN	+	+	+	+
QoS: traffic separation	-	+	+	+
PFC	-	-	partial	-
Performance	←			
Ease of Config	←			



- 不同的应用有不同的通讯模式
- 网络对于突发Burst的吸收问题
- 通讯中的流量分配问题
- 如何降低网络转发的延时问题
  - Cut Through & Store Forward

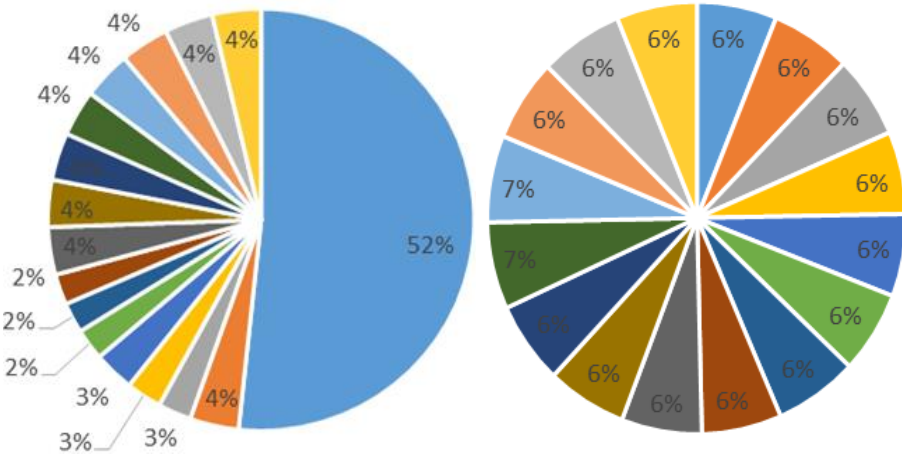
○ ○ ○ ○ ○ ○



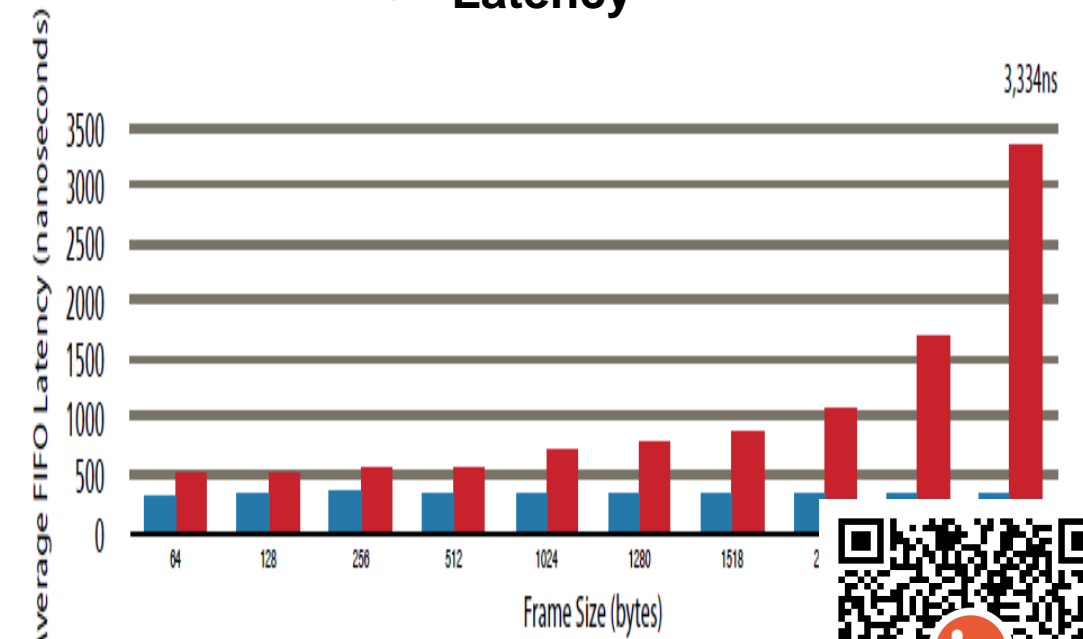
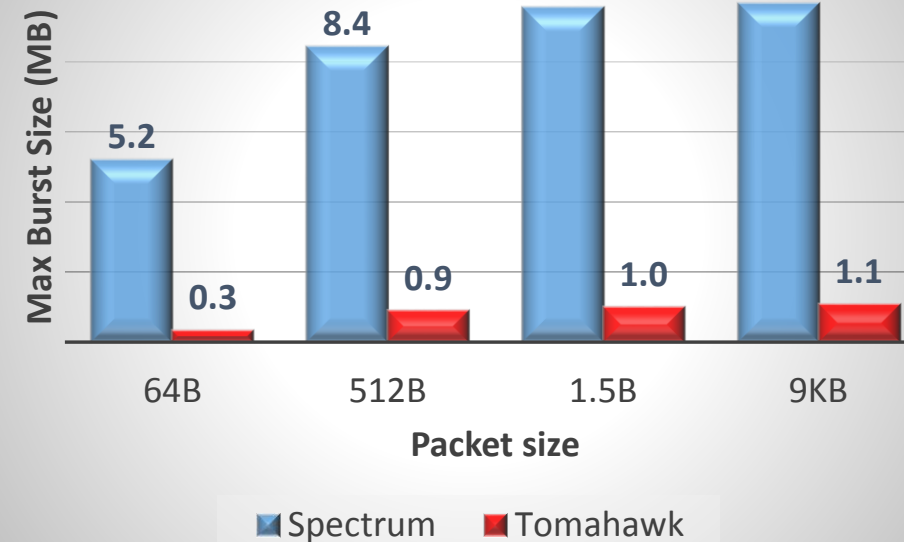
Many to One Latency

Broadcom

Spectrum



Microburst Absorption Capability



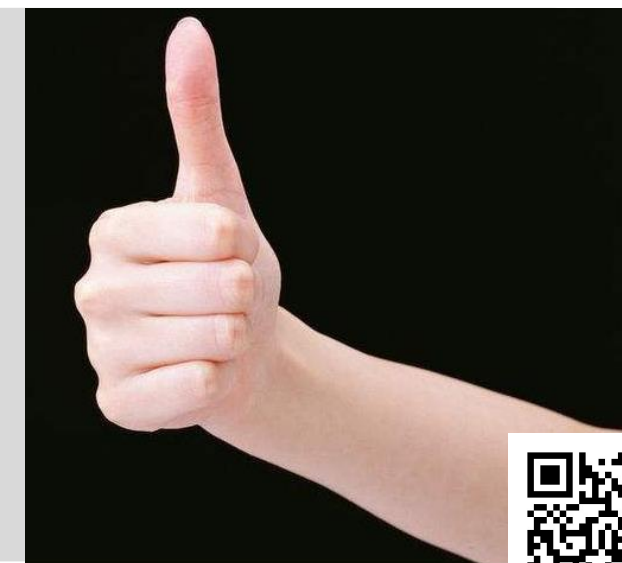
微信扫码收听演讲音频



数据中心标准化、简单化、高性能化



应用和网络相结合，网络为应用服务





Thank You

