Deterministic IP

刘冰洋，华为网络技术实验室主任研究工程师
liubingyang@huawei.com

清华大学“计算未来”博硕论坛023期特邀报告
2018.4.22

背景：确定性网络传输是业界的热门话题

- IEEE Time Sensitive Networking (TSN)
  - 工业控制网络、车载网络等对传输的可靠性、**延时抖动**的确定性有极高要求
  - 尤其对于抖动，不同业务的要求端到端在~10us或~100us级别
  - IEEE TSN制定基于以太网标准的确定性技术，收编七国八制的零散技术，受到业界的极大关注，也成为通信设备公司进入工业网络市场的切入点
  - 但TSN标准局限在局部二层网络，范围和规模严重受限
- IETF Deterministic Networking (DetNet)
  - 确定性需求绝不限于局部二层网络，来自不同机构的二十多位作者联合撰写了确定性网络Use Case文稿，阐述了在九大产业里的需求，包括pro audio&video, electrical utilities, building automation systems, wireless for industrial, cellular radio, industrial M2M, mining industry, private blockchain and network slicing等
  - DetNet在三层网络解决以上需求，在统计复用的基础上提供确定性时延和抖动。DetNet成为IETF最受关注的工作组之一，架构明确后将提案数据面技术和标准

Deterministic Latency is Required in Large-scale, Layer-3 Networks

- IETF DetNet WG focuses on deterministic layer-3 data paths
  - DetNet scope includes very **large networks**, e.g., Utility Grid network, spanning a whole country, and involving many hops
  - Applications and IT services are transitioning to IP. **TSN can't satisfy the needs**.
- Strong requirement on **deterministic low latency and bounded jitter**

1. Electrical Utilities - Teleprotection systems ideally support zero asymmetric delay; typical legacy relays can **tolerate delay discrepancies of up to 750us**"
2. Building Automation Systems (right figure) – **End-to-end jitter should be less than 1ms**
3. Cellular Radio – The "midhaul latency" and "channel state information" reporting among CoMPs is **delay-sensitive limited in 1ms ~ 10ms**.
4. Industrials M2M – requires converged IP-standards-based Network with **bounded latency and jitter**.

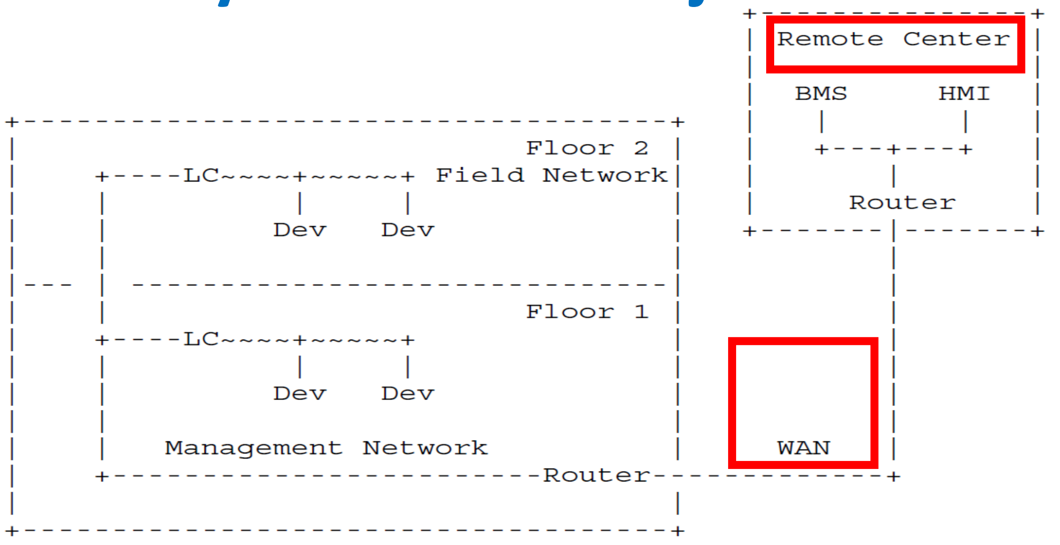
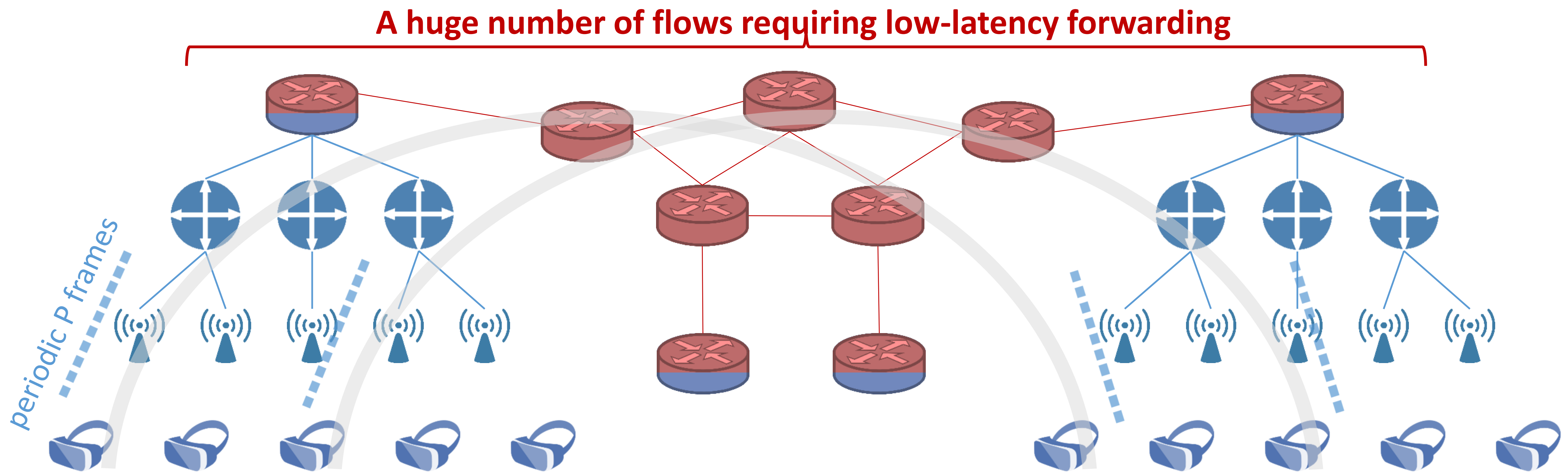


Figure 6: Deployment model for Small Buildings

An Example: VR Real-time Communication



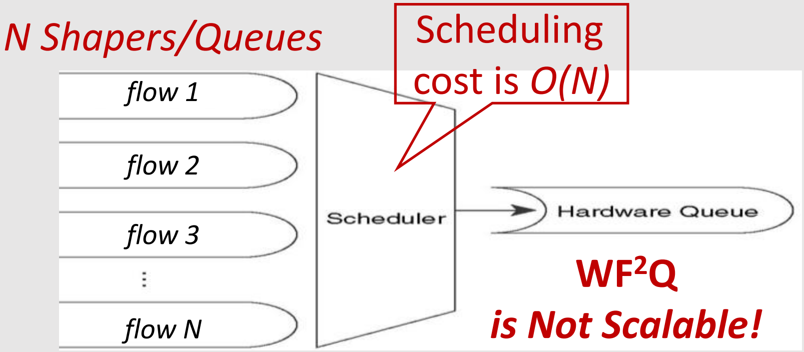
VR real-time communication requirement on latency:

- Application-layer end-to-end latency <= 20ms. This includes motion capture, rendering, etc.
- Latency budget for network transfer is only 5ms ~ 7ms, including air interface (2ms) and propagation delay
- If link propagation delay is 2ms (400km), the budget for **end-to-end queuing delay is only 1ms ~ 3ms**
- If there are 10 hops, **per-hop queuing delay budget is only 100us!**

No Technique can Simultaneously Achieve **Deterministic Latency** and **Scalability**

	Deterministic Latency	**Scalability**
Earliest deadline first scheduling	**No** *Jitter accumulate linearly*	**No** *Packet scheduling cost is not O(1)*
TSN 802.1Qbv - time slot based scheduling	**Yes**	**No** *Slot assignment is NP hard*
Class based priority queue (Even if per-flow shaping at ingress edge)	**No** *Jitter accumulate linearly*	**Yes**
Per-flow shaping & scheduling	**Yes**	**No** *Per-flow queue. Packet scheduling cost*

For example, Worst-case Fair Weighted Fair Queuing (WF^2Q) achieves deterministic queuing latency, but it is not scalable.



DIP for **Large-scale Deterministic** Networks

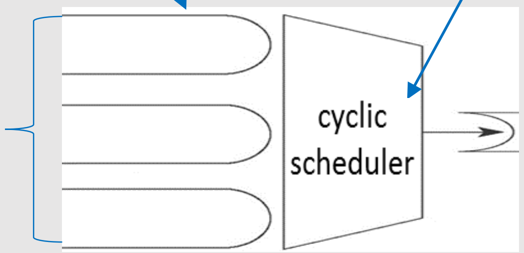
- **DIP achieves both scalability and end-to-end deterministic latency** with *core-stateless cyclic queuing and scheduling*
- Preview of the key features (e.g., T=25us):

✓ Scalable DIP core routers

Only 3 deterministic queues per interface.
Coexist with BE queues

Per-packet scheduling cost is O(1)

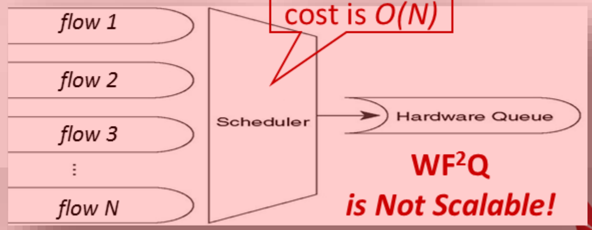
Buffer size is only 938KB for 100G interface



Low cost!

N Shapers/Queues

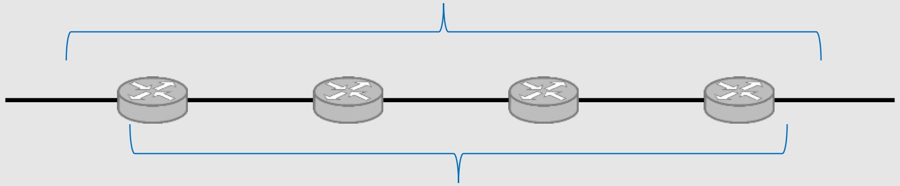
Scheduling cost is *O(N)*



WF^{2}Q is Not Scalable!

✓ End-to-end deterministic latency

End-to-end jitter <= 50us, irrelevant to number of hops or distance

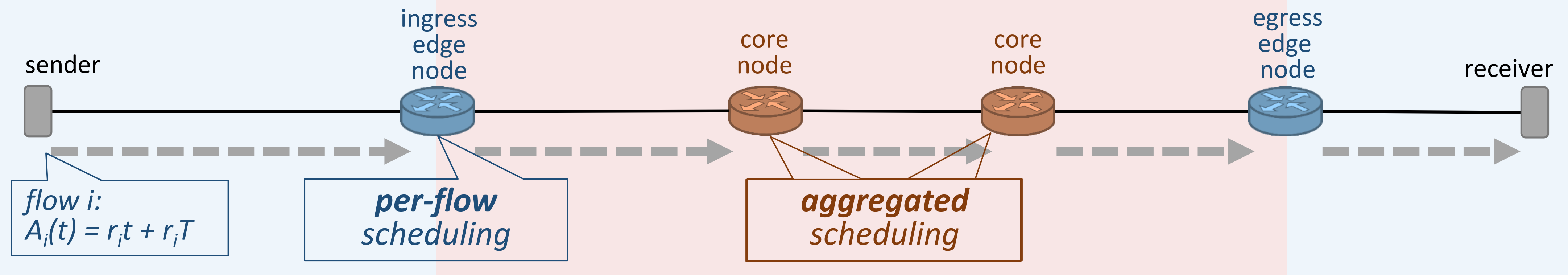


Per-hop expected queuing delay is 1.5T = 37.5us << 100us.
H-hop expected queuing delay is 1.5hT.
If h=10, expected queuing delay is 375us << 1ms

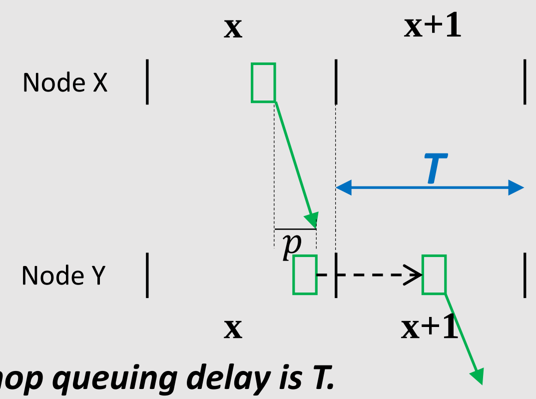
New service!

Overview: **Core-stateless Framework**

- Ingress edge nodes perform per-flow cyclic queuing and scheduling
- Core nodes are stateless, perform aggregated queuing and scheduling



✓ Cyclic Queue Scheduling



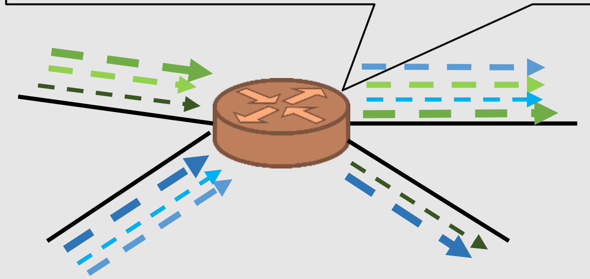
Per-hop queuing delay is T.
End-to-end h-hop queuing delay is hT. Jitter is 2T.

Key Principle

- Hop-by-hop shaping to mince bursts into *micro burst cycles*
- A micro burst cycle is short. E.g., cycle duration T = 10us
- A micro burst is very small. Burst size ≤ TR. $R = \Sigma r_i$
- Packets in a same cycle are in a same queue. No flow queue

✓ Aggregated States

Control Plane: $\mathbf{R = \Sigma r_i \le C}$
Data Plane: a micro burst **≤ TR**



HUAWEI

Theoretical Evaluation

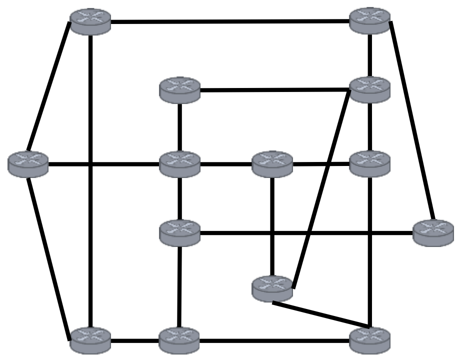
- DIP provides **deterministic end-to-end latency**
  - End-to-end jitter <= **2T**
  - End-to-end queuing delay = **$\Sigma(T+\tau)$**
    - For *h* hops, expected delay is **1.5hT**
- DIP is very **scalable on core nodes**
  - Per-interface only **3 queues**
  - Per-interface buffer size is **3TC**
  - Per-packet scheduling cost is **O(1)**

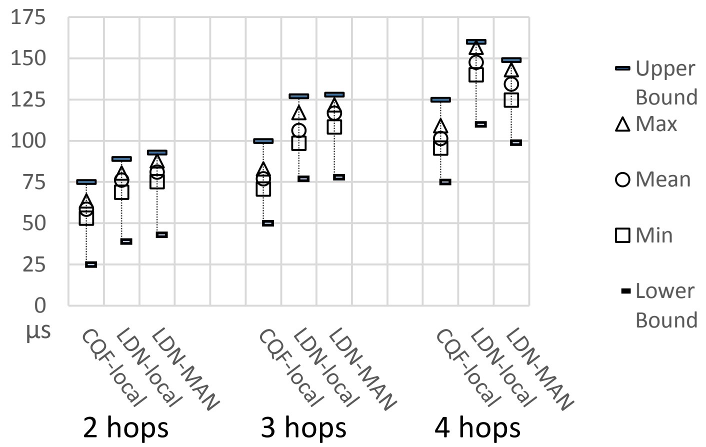
For instance, if T=25us

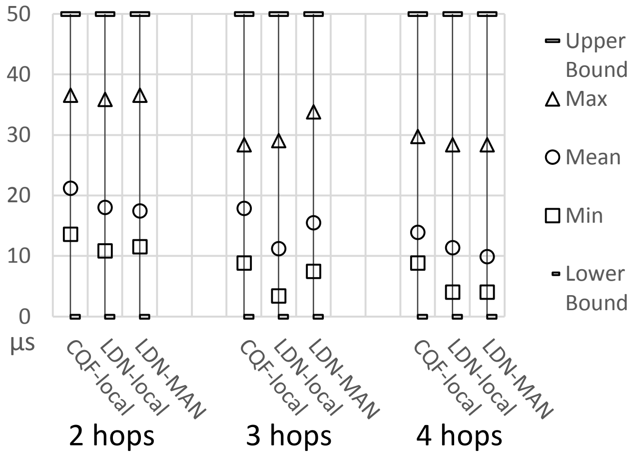
- End-to-end **jitter <= 50us**
- End-to-end queuing expected **queuing delay is 375us**, if h=10
- Per-interface buffer size = **938KB** if link speed is 100Gbps

Simulation Results

- We simulate DIP in 2 scenarios: local network and MAN. T = 25us, C=1Gbps
  - In the local network scenario, we compare with CQF.

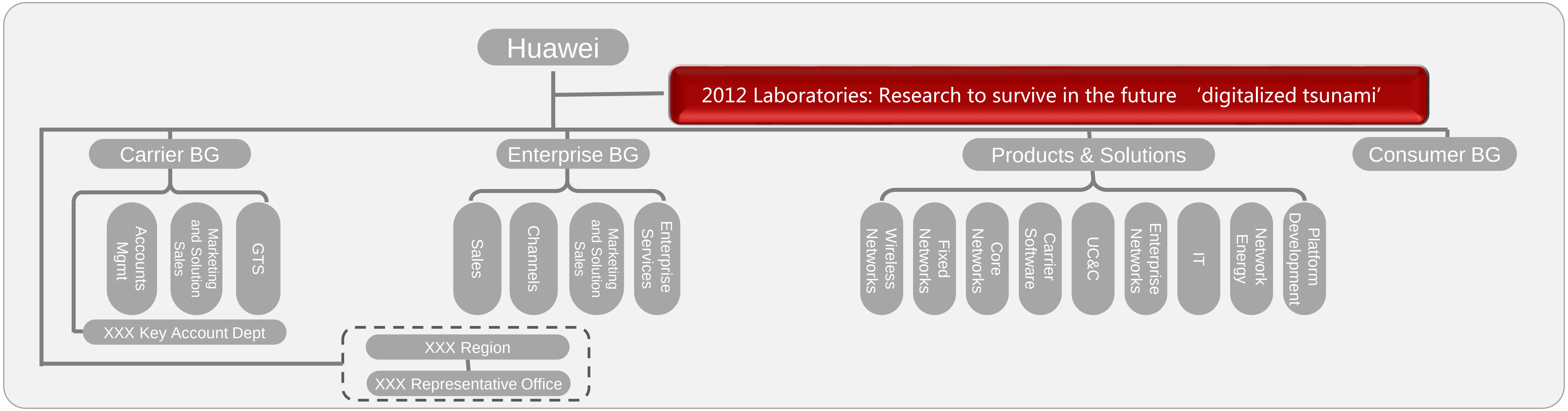


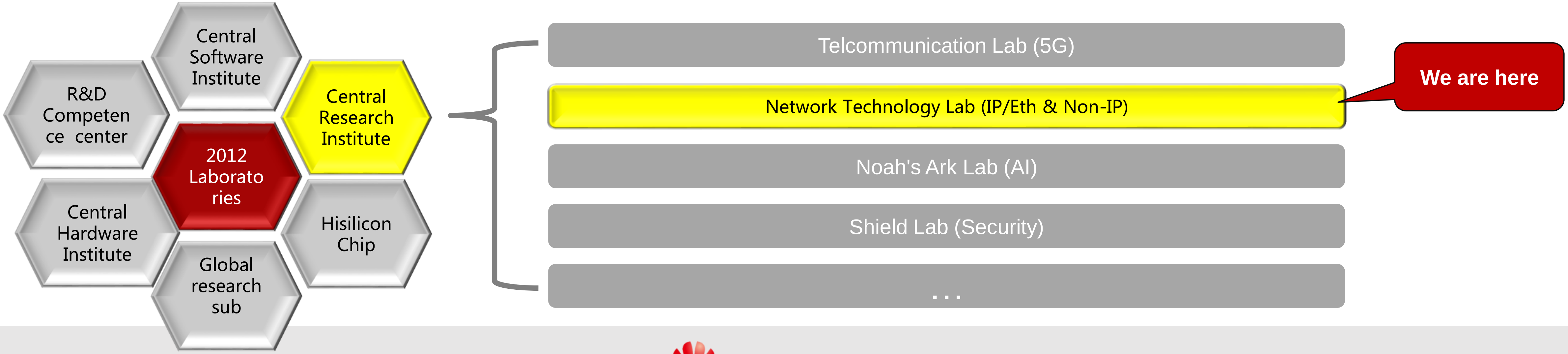




- The simulation results perfectly follow theoretical analysis
  - Jitter is constant. Queuing latency grows “linearly”. Buffer bound.

网络技术实验室在华为公司组织架构中的位置





网络5.0的研究方向

构建太平洋粗管道

- 分布式光互联路由器PRouter
- 超大带宽数据中心网络NG-DCN

网络协议的下一跳

- IP下一跳/New IP
- 下一代以太网：FlexE/X-Ethernet

机制上解决低时延问题

- X-E：业界首创恒定百纳秒级低时延交换
- 无损网络：高性能计算、人工智能呼唤无损网络

拓展新空间

- 工业4.0驱动下的产业互联网
- 创新的极简企业网架构
- 互联网第二平面：体验驱动的overlay云网络

欢迎加盟网络技术实验室

关于招聘：

- 面向2019年毕业的博士研究生。
- 网络技术实验室的详细介绍以及招聘岗位信息，请扫描如下二维码。
- 重点招聘方向：协议体系研究、芯片架构、编译算法、路由器架构、以太网、数据中心网络、产业互联网/IOT网络、自组织网络等。

关于实习：

- 面向目前在读的硕士和博士研究生；
- 博士实习生：实习工资10000/月；本/硕实习生：实习工资6000/月；
- 在华为实习期内，公司为每位同学购买了人身意外保险；
- 工作安排：深入了解华为，和资深华为专家一起探索新技术、新想法，参与技术讨论、方针、验证；
- 实习周期建议再3个月以上；

- **2017年，华为北京研究所应届博士生的北京户口全部得到解决。**
- **实验室博士65人(30%)，其中清华博士6人、博后1人、副教授一人。**
- **华为北京研究所正在筹建博士后工作站，争取为2019届博士提供更多选择。**

Thanks!