XILINX

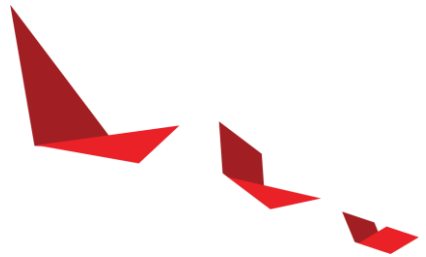
基于智能网卡的HyperLedger Fabric硬件加速器

Accelerating Hyperledger Fabric within SmartNIC

杨骥 胡成臣

赛灵思实验室

目录



- 赛灵思 (Xilinx) 实验室简介
- HyperLedger Fabric的性能瓶颈
- Blockchain Machine: 基于智能网卡的Fabric加速方案
- 性能评估
- 开源资料

- 致谢：Haris Javaid, Nathania Santoso, Mohit Upadhyay, S Mohan, Chengchen Hu, Gordon Brebner

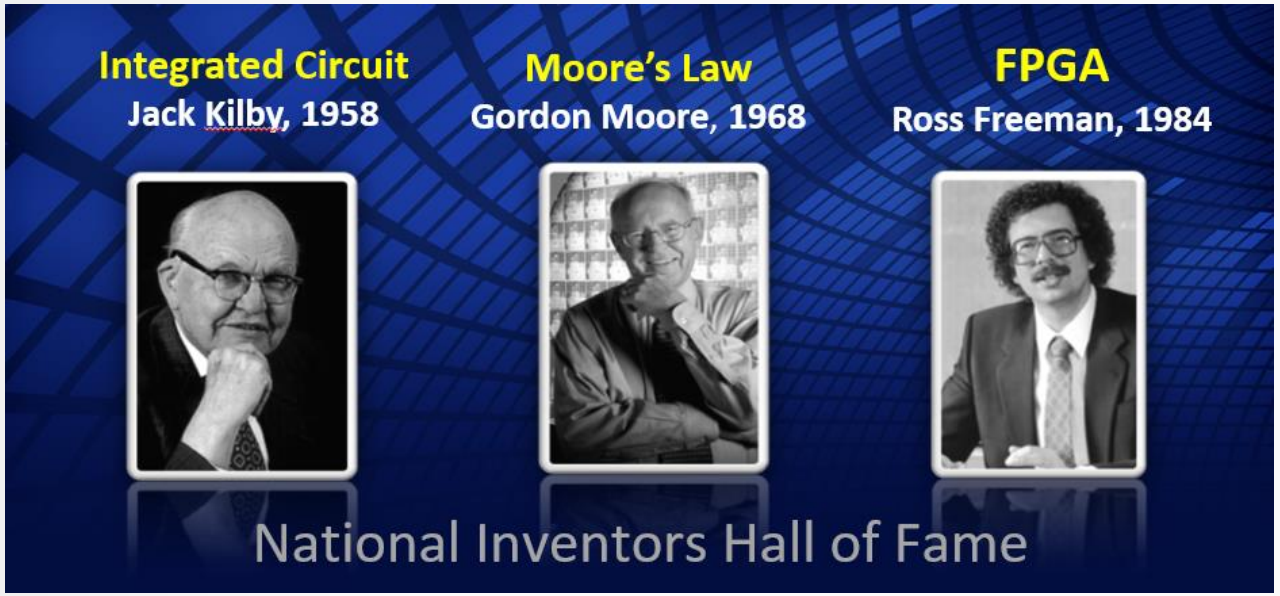
赛灵思简介

Xilinx Founded in 1984, Go Public in 1990

6th Largest Fabless Semiconductor Corp.

FY21 revenue	**$3.15B**
market segment share	**~60%**
employees worldwide	**~5,000**
customers worldwide	**20,000**
patents	**3,500+**
industry firsts	**60+**

Inventor of the FPGA
National Inventor's Hall of Fame



XILINX.

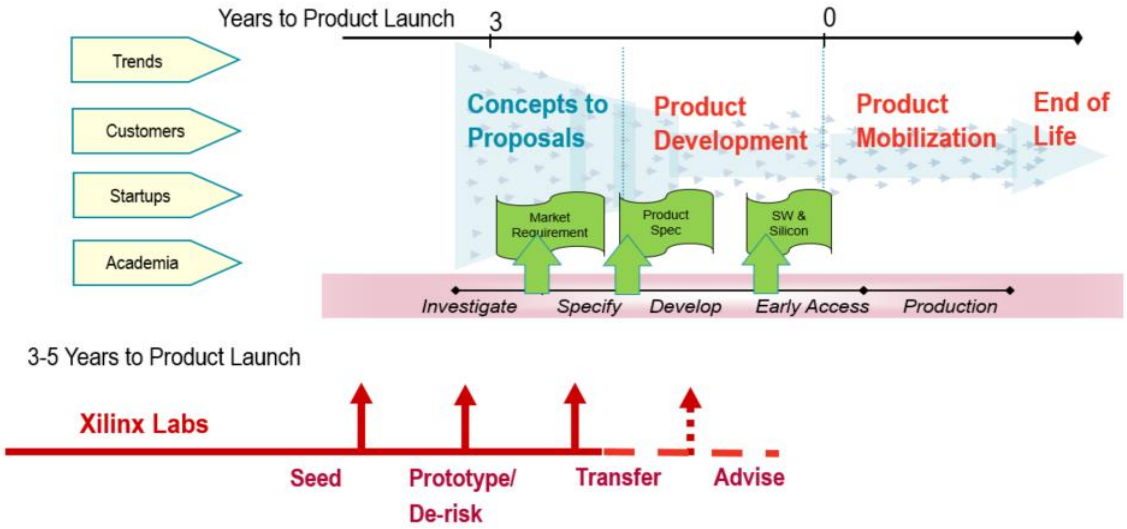
赛灵思实验室介绍

Xilinx CTO Organization

- Toward technologies of tomorrow
  - Explore, innovate, differentiate, derisk and deliver to products
  - Enable new business/users and seed new market opportunities
  - Provide a ‘more than Moore’ roadmap
  - Win the mindshare of start-up and research communities

Locations

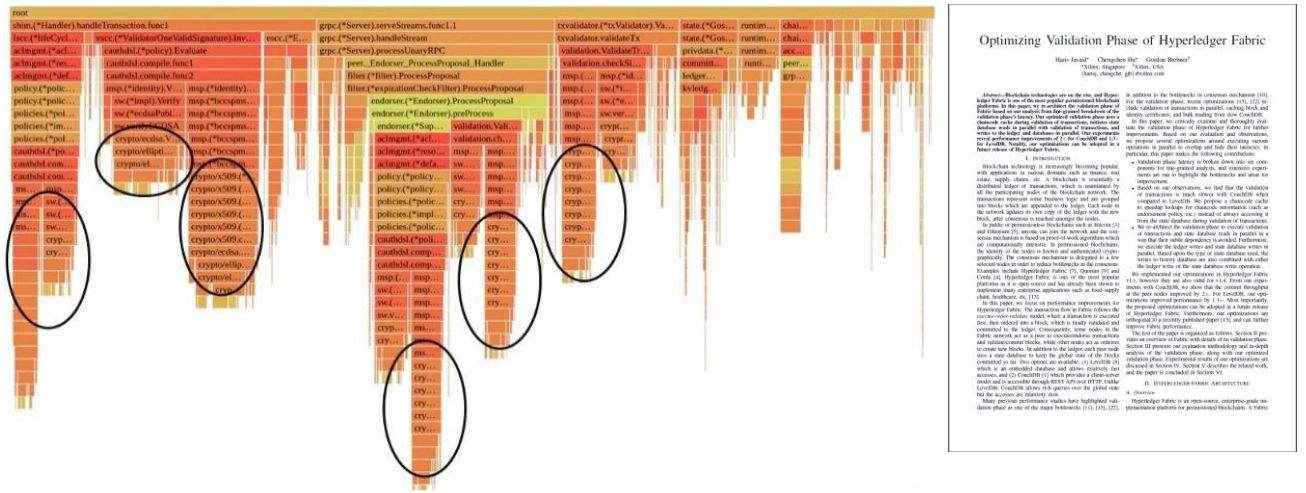
- **San Jose, California, USA (Xilinx North America HQ)**
- Longmont, Colorado, USA
- Dublin, Ireland (Xilinx Europe, Middle East, Africa HQ)
- **Singapore (Xilinx Asia Pacific HQ)**



区块链相关工作

- 公链研究
  - 基于Stratum矿池代理的挖矿加速方法
  - ANT Pool 实验矿池部署
  - 会议论文Crypto Valley 2019

- 联盟链研究
  - HyperLedger Fabric性能分析
  - 主要性能瓶颈来自验证(Validation)阶段
  - 会议论文IEEE MASCOTS 2019

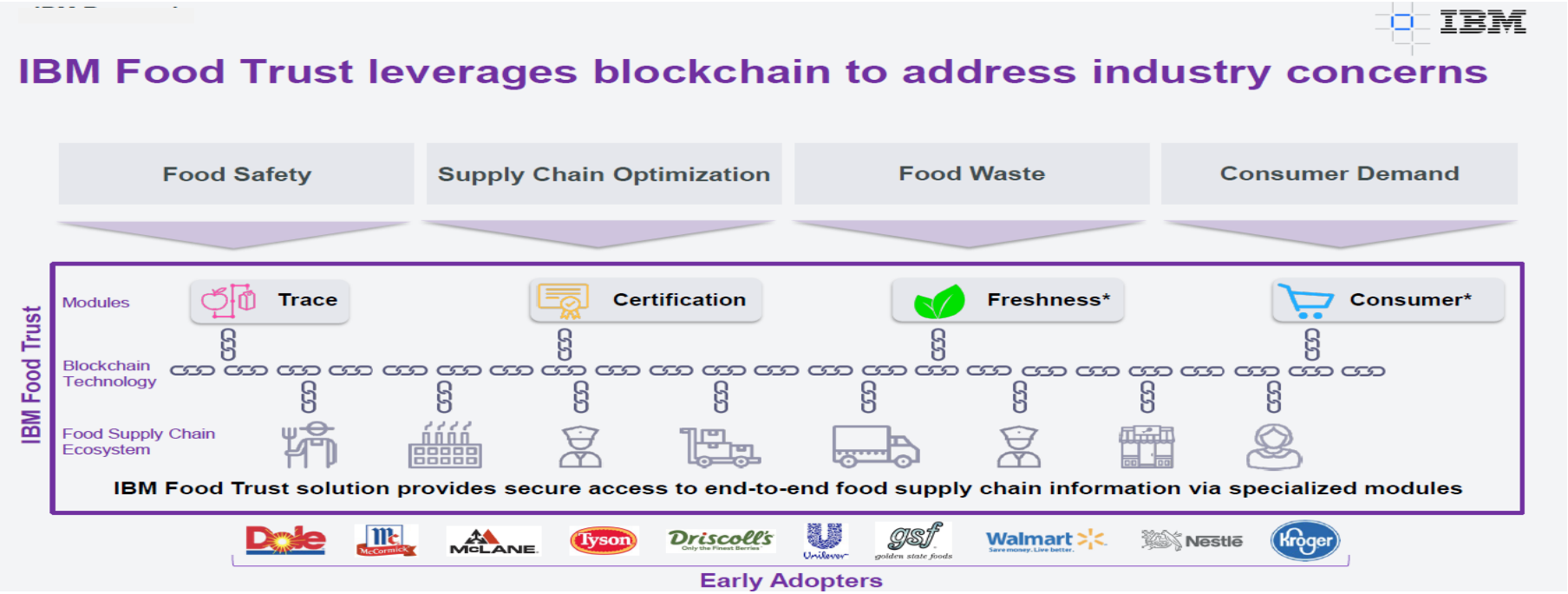


XILINX.

HyperLedger Fabric 可扩展性问题

- Fabric吞吐率相比公链有很大提升
  - 原始Fabric ： ~600 tps
  - 优化执行：~7700 tps
  - 优化实现+优化网络：~22000 tps
  - vs. Bitcoin: ~5 tps
  - vs. Ethereum: ~15 tps
  - vs. Ethereum 2.0: ~10000 tps (expectation)
- 与Visa等交易系统相比仍然差距较大
  - Visa 65000 tps

HyperLedger Fabric 可扩展性问题



- Example scaling up:
  1. One retailer, selected products
  2. One retailer, all products
  3. Major retailers, all products
  4. All retailers, all products

> Estimated transactions per second:
  1. 40,000
  2. 400,000
  3. 2,500,000
  4. 250,000,000

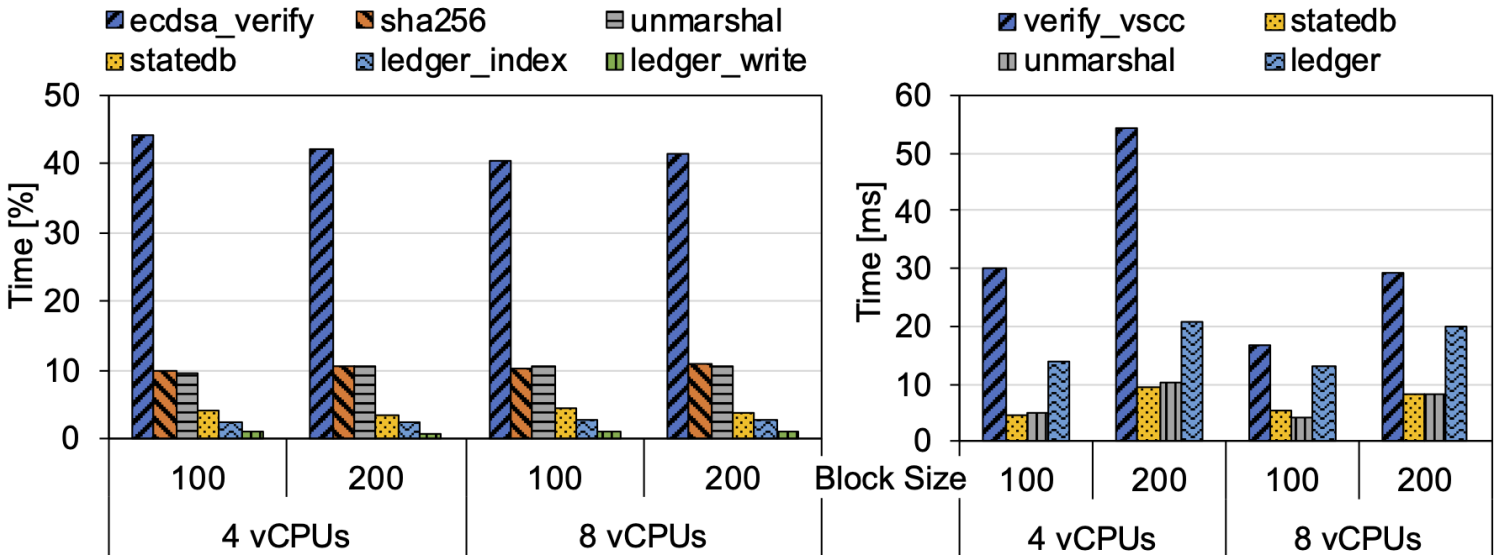
Software-only wall: 400K, based on today's best plus 10x future optimization

Another 10x-100x needed from acceleration

XILINX

HyperLedger Fabric 的主要瓶颈

- 验证是Fabric中的主要瓶颈[1,2,3]
  - 数据通信中涉及大量protobuf序列化反序列化
  - 验证过程中的ECDSA签名验证需要大量计算资源
  - 状态数据库（stateDB）访问速度较慢
  - 块较大时Ledger写入速度慢，且受到I/O速度制约

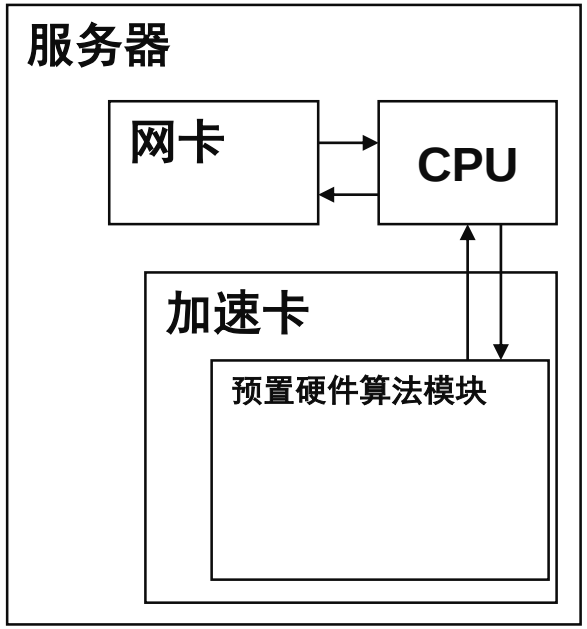


[1] P. Thakkar, S. Nathan, and B. Vishwanathan, “Performance Benchmarking and Optimizing Hyperledger Fabric Blockchain Platform,” in *MASCOTS*, 2018.
[2] C. Gorenflo, S. Lee, L. Golab, and S. Keshav, “FastFabric: Scaling Hyperledger Fabric to 20,000 Transactions per Second,” in *ICBC*, 2019.
[3] P. Thakkar and S. Nathan. 2021. Scaling Hyperledger Fabric Using Pipelined Execution and Sparse Peers. arXiv:2003.05113.

XILINX.

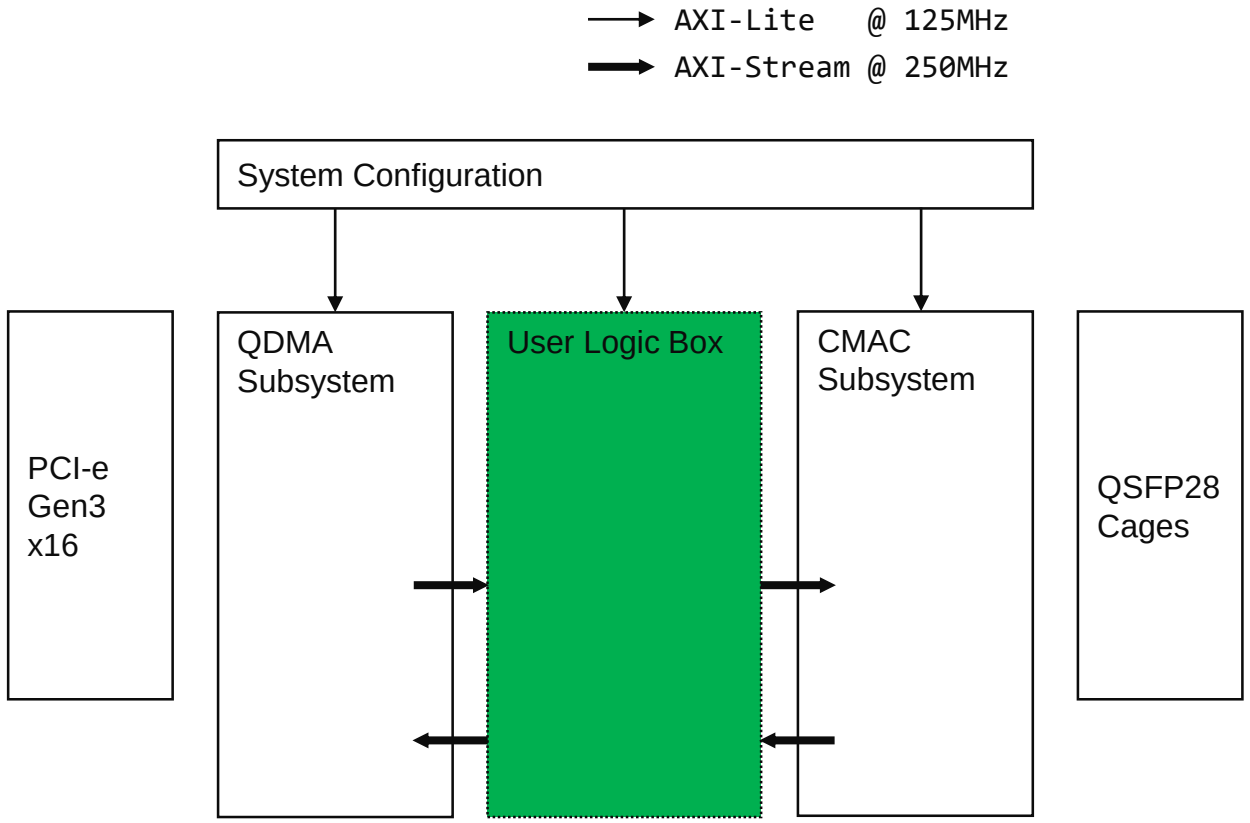
HyperLedger Fabric 硬件加速方案局限性

- 主流方案：PCIe密码学算法加速卡
  - 算法局限：只对特定算法进行加速
  - 带宽浪费：PCIe，主内存消耗
    - 待加密（解密）数据通过PCIe搬移到加速卡
    - 加密（解密）后数据通过PCIe搬移回主存储



Xilinx Labs OpenNIC

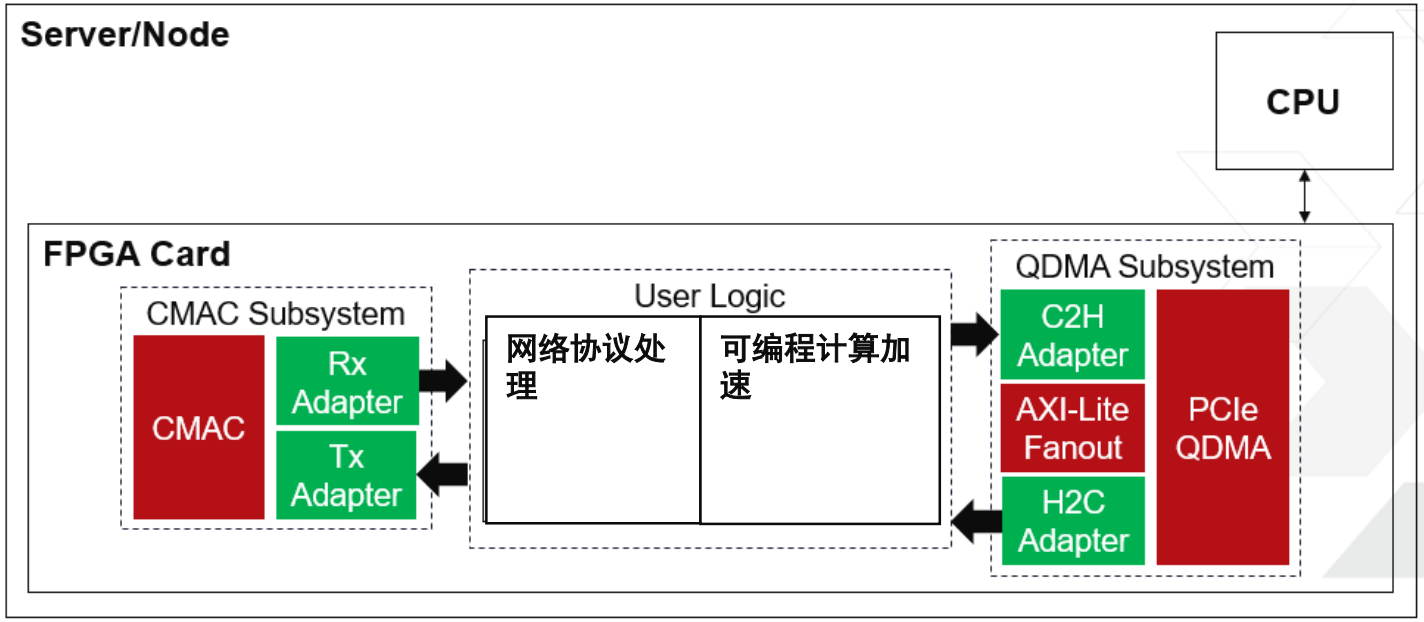
- **CMAC subsystem**
  - Support up to 2 ports (QSFP28)
- **QDMA subsystem**
  - Support up to 4 physical functions
  - Support up to 2048 queues
- **Data path**
  - AXI-stream with 512b data
  - Run at 250MHz
- **Control path**
  - AXI-lite with 32b address and data
  - Run at 125MHz, phase-aligned with the 250MHz AXI-stream
- **Linux Kernel Driver**



基于OpenNIC的计算加速模型

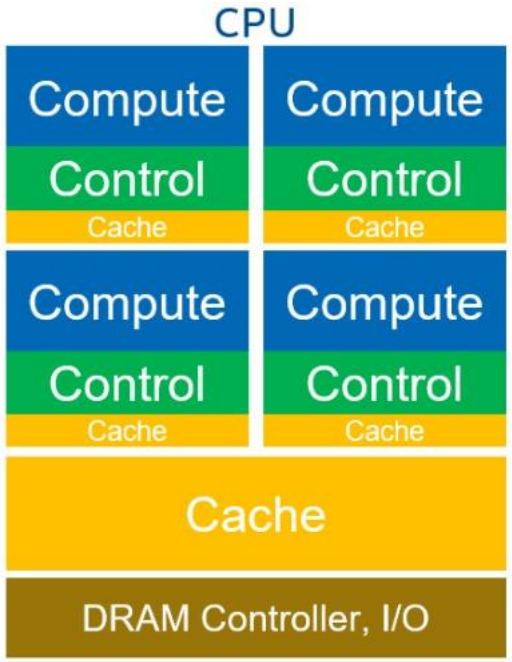
- 智能网卡方案
  - 瓶颈计算任务卸载到智能网卡
  - 数据流经智能网卡时完成计算加速**智能网卡**

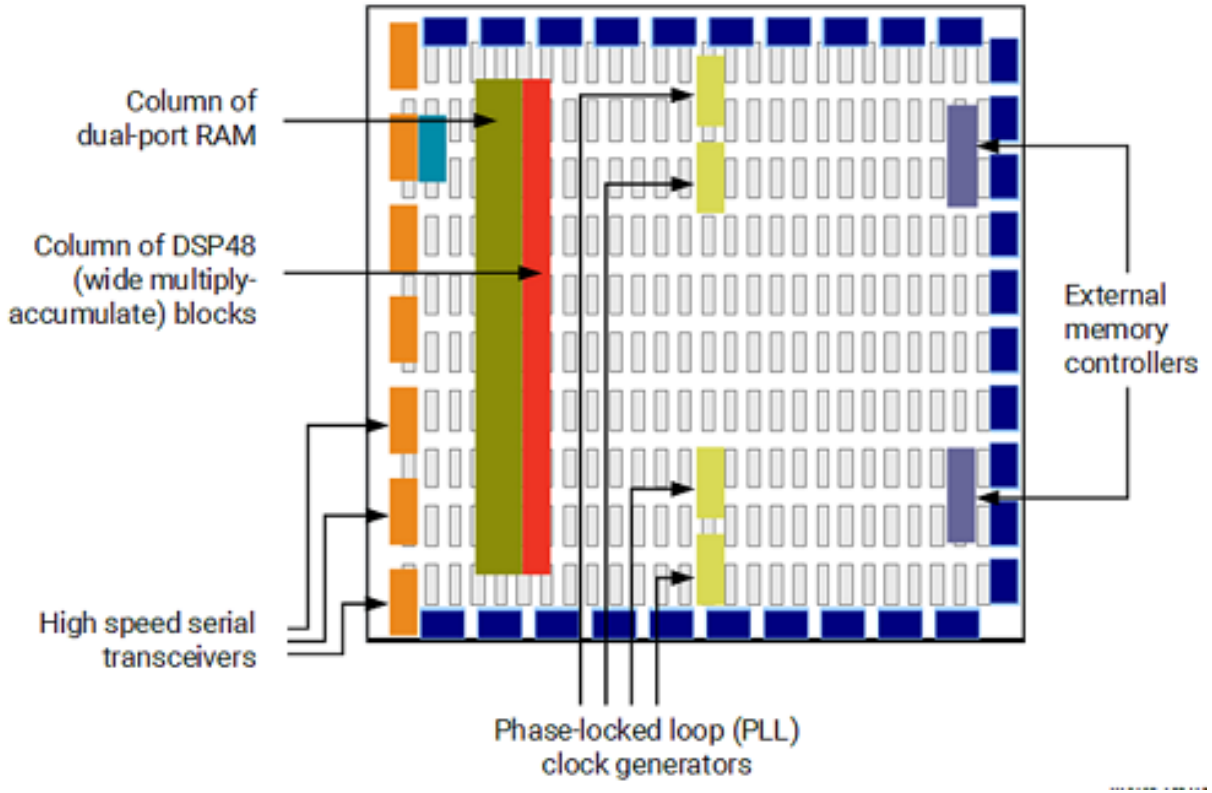
Active Option



XILINX

CPU, FPGA 比较

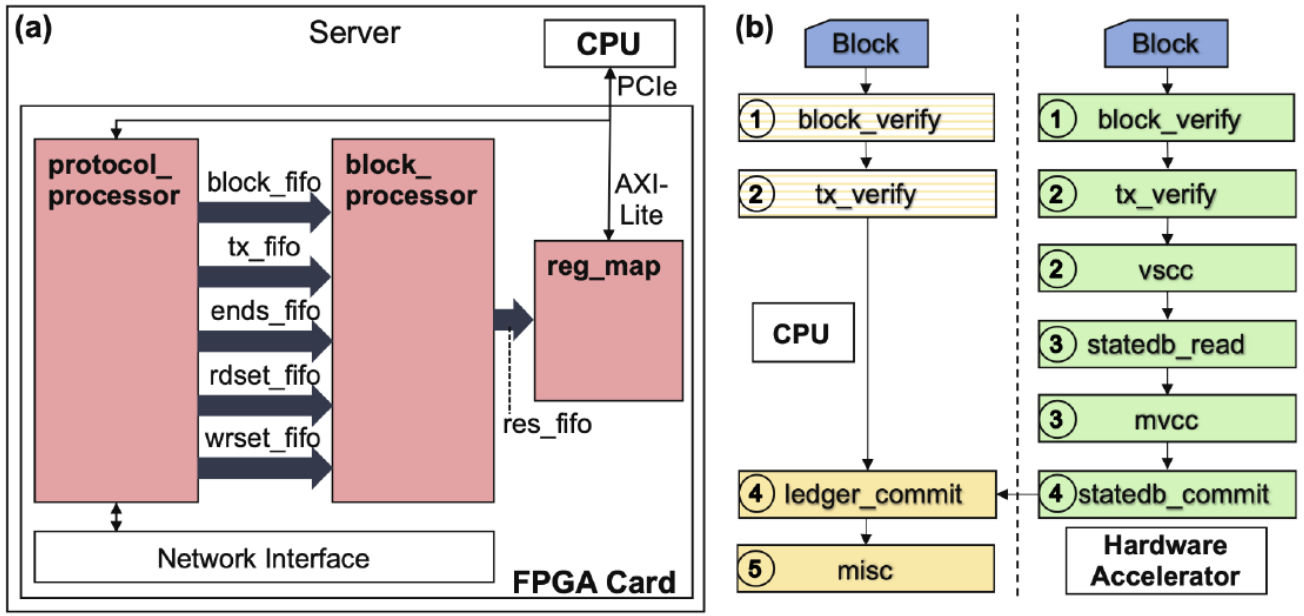




Reference: [1] Compare Benefits of CPUs, GPUs, and FPGAs for Different oneAPI Compute Workloads
[2] Understanding FPGA Architecture

HyperLedger Fabric 加速方案 ---- Blockchain Machine

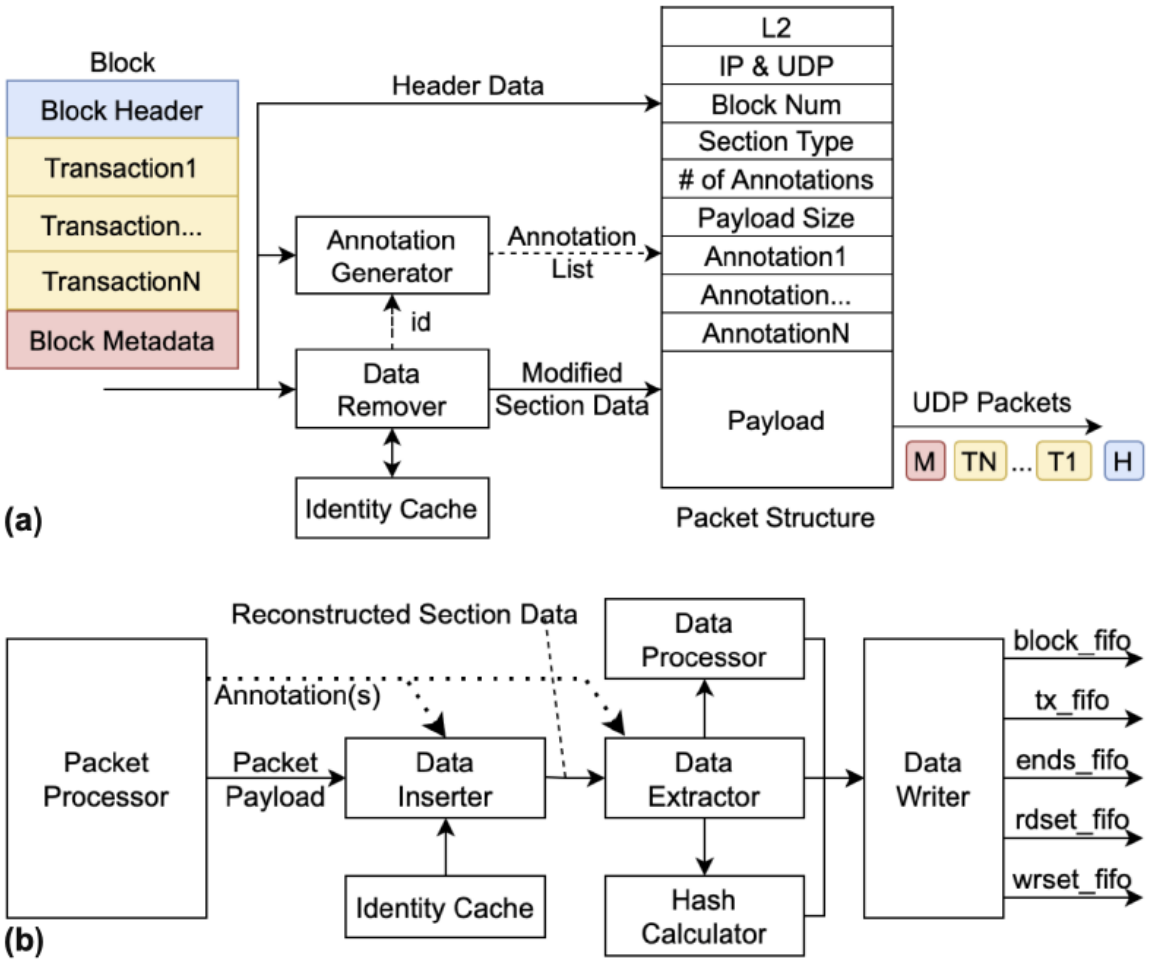
- 系统结构与软硬件协作分工
  - （a）加速方案硬件结构
  - （b）软硬件协作分工
    - FPGA硬件负责执行计算相关加速
      - VSCC
      - State DB Read
      - MVCC
    - 服务器CPU执行非瓶颈任务
  - 数据流
    - Protocol Processor 首先处理块，并提取数据组成加速任务交给后级
    - Block Processor 包含块与交易两组流水线，并行处理加速任务，包括ECDSA验证，状态数据库，寄存器等
    - Reg Map将处理后的块以及交易信息提交到Ledger



XILINX.

HyperLedger Fabric 加速方案

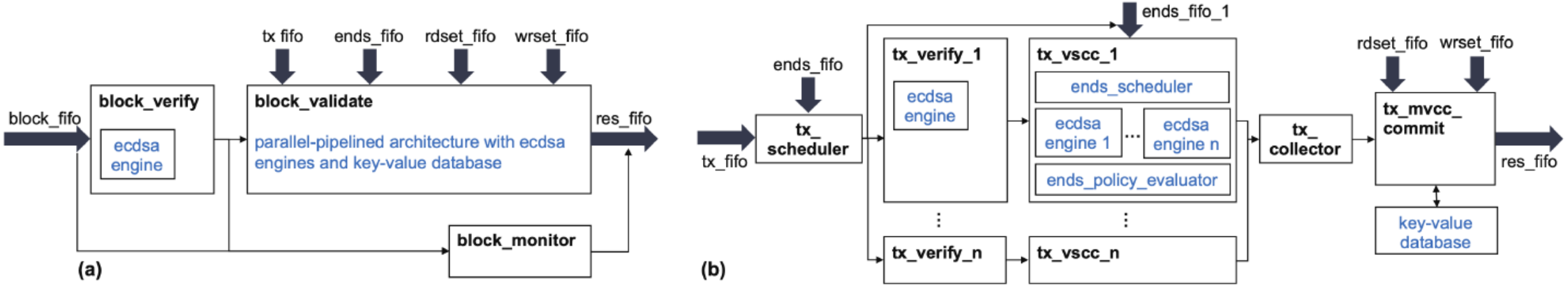
- Fabric网络协议存在的性能牺牲
  - 基于gRPC/HTTP2的fabric网络传输协议
    - 消耗CPU需要迭代解析对硬件不友好
    - 包含跨区块的重复信息（Identities）
- 加速方案：基于UDP的自定义协议
  - 使用自定义UDP传输协议
  - 标记需要硬件加速的数据
  - 引入缓存，删除已经被缓存的跨块重复信息
  - （a）自定义协议发送端数据流
  - （b）自定义协议硬件处理器模块



XILINX.

HyperLedger Fabric 加速方案

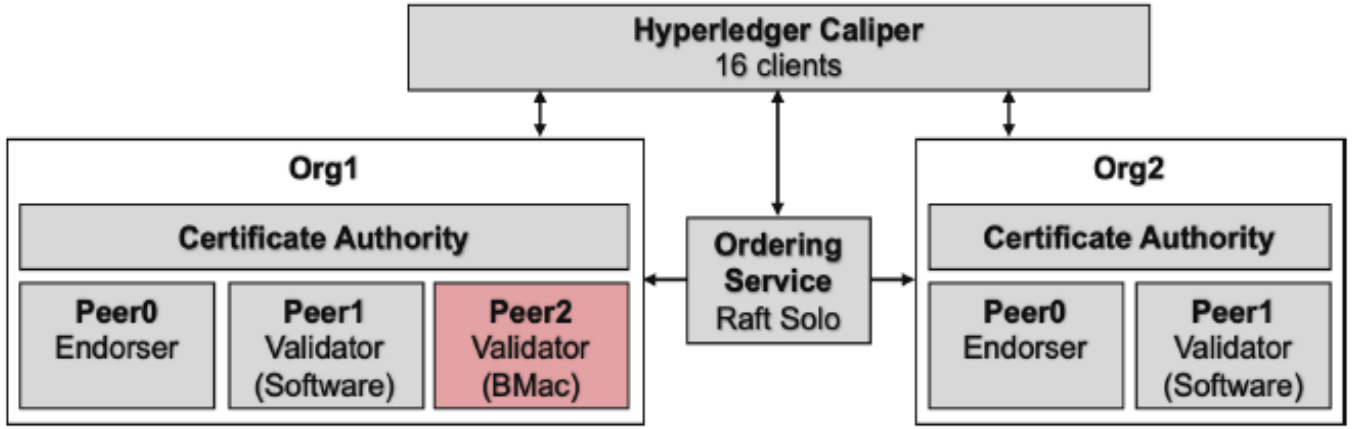
- 块数据处理模块
  - 硬件状态数据库
  - 块签名验证
  - 交易签名验证
    - 交易并行化处理



XILINX.

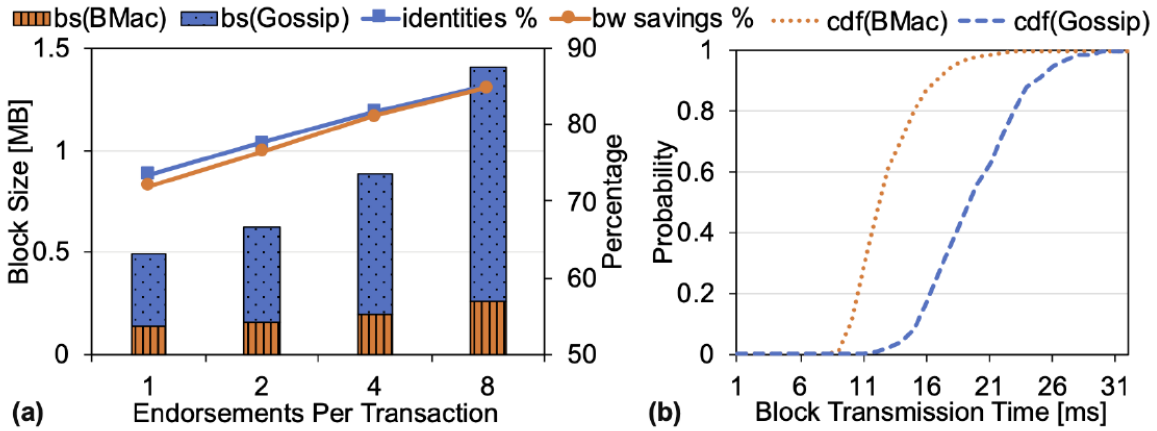
性能测试

- 实验平台
  - Xilinx Alveo U250板卡，加载OpenNIC
  - 拓扑: 单Orderer，2 Organizations， 2 Peers
  - Caliper：smallbank，DRM
  - 服务器配置：Xeon 2.2G，2GB RAM/vCPU
  - 100G 网络
  - 对比节点计算任务线程使用独立vCPU



性能测试

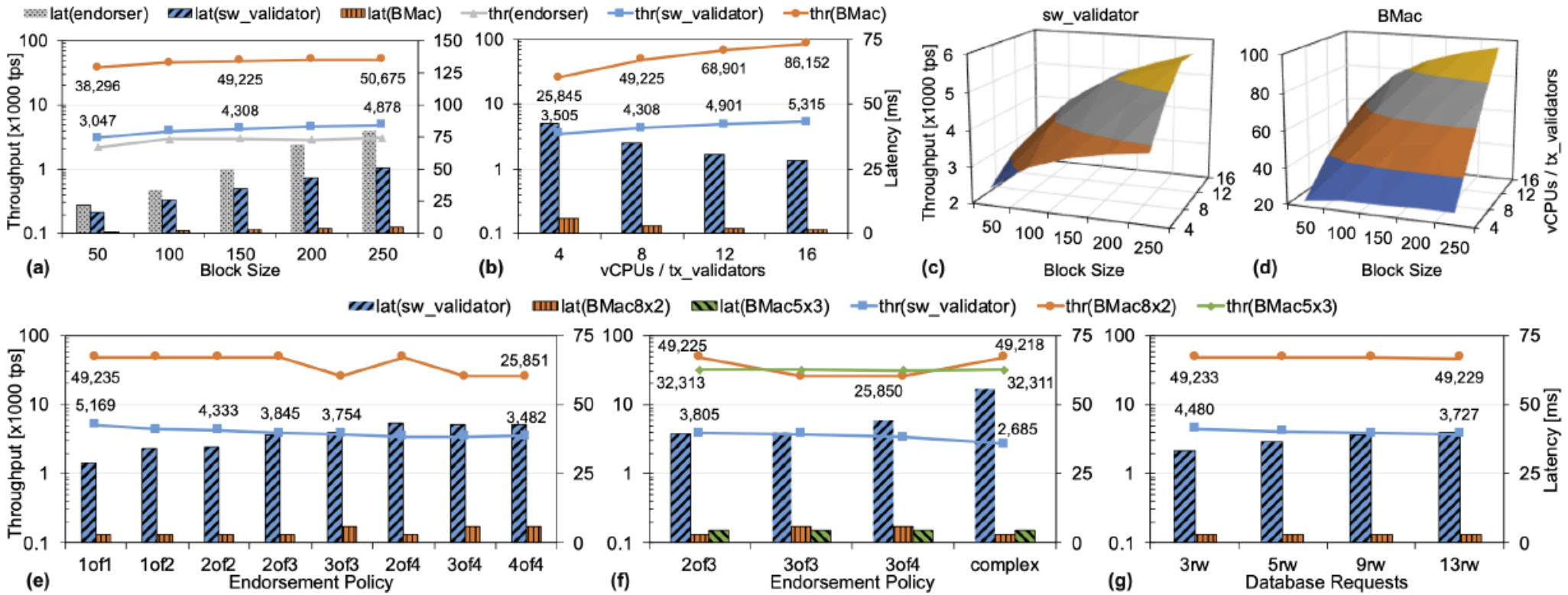
- 网络协议性能
  - （a）数据压缩：3.4x ~5.3x
  - （b）带宽节约85%
  - （c）块传输尾部时延降低30%
  - （d）处理能力~1M tps (使用发包软件测试）



Endorsements Per Transaction	Packet Size [Bytes]	Max. Processing Rate [Gbps]	[tps]
1	921	10.6	1.14M
2	1002	11.1	1.11M
3	1080	11.4	1.05M
4	1164	11.6	996K

性能测试

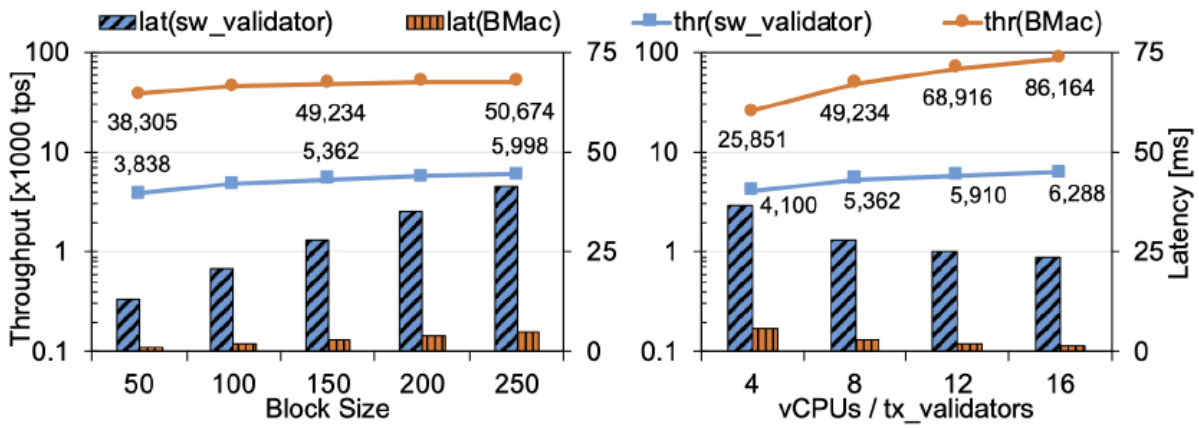
- 端到端性能测试
  - Smallbank在不同配置下数据对比
  - 最佳情况下提交吞吐率为95600 tps,验证时延5ms （**4.5x** 相比于目前有报道的最佳性能）



XILINX

性能测试

- 端到端性能测试
  - DRM在不同配置下数据与Smallbank相似



XILINX.

资源占用率

- 基于Alveo U250 FPGA加速卡的资源占用率

Resource	4x2	8x2	12x2	16x2
LUT / LUTRAM	20.9%	28.5%	35.8%	43.3%
FF	6.9%	8.0%	9.1%	10.3%
BRAM / URAM	13.1%	13.1%	13.1%	13.1%

- * 4x2表示：共有4个并行tx验证核心，每个vscc中使用两个ECDSA加速器

更多细节

- 1. Blockchain Machine论文
  - https://arxiv.org/abs/2104.06968
- 2. 开源代码 HyperLedger Labs
  - 即将上线，敬请期待
- 3. 赛灵思自适应计算研究集群(Xilinx Adaptive Compute Clusters)
  - NUS节点即将开放
- 4. OpenNIC 开源代码
  - OpenNIC umbrella repo: https://github.com/Xilinx/open-nic
    - Apache v2 license

Thank You