Direct Universal Access: Making Data Center Resources Available to FPGA

Ran Shu[1], Peng Cheng[1], Guo Chen[2],
Zhiyuan Guo[1, 3], Lei Qu[1], Yongqiang Xiong[1],
Derek Chiou[4], Thomas Moscibroda[4]

Microsoft Research[1], Hunan University[2],
Beihang University[3], Microsoft Azure[4]

Microsoft

FPGA Deployment in Data Centers

- Wide deployment
  - Major cloud service providers
    - Microsoft, Amazon, Facebook, Alibaba, Tencent, Baidu, IBM, etc.
- Accelerated applications
  - Computation
    - Web search ranking
    - Deep neural networks
    - Big data analytics
  - Networking
    - Network processing
  - Database/Storage
    - SQL
    - Key-value store

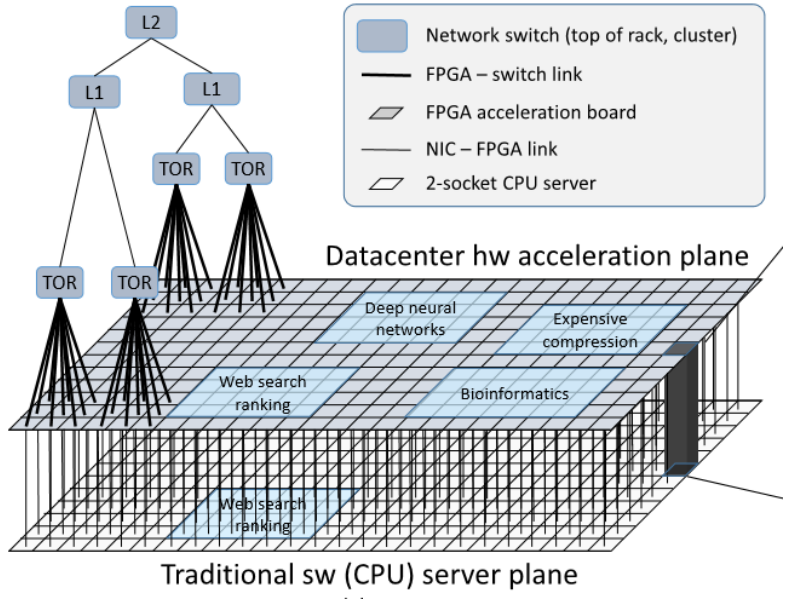


Image from A. Caulfield et al., Micro 2016

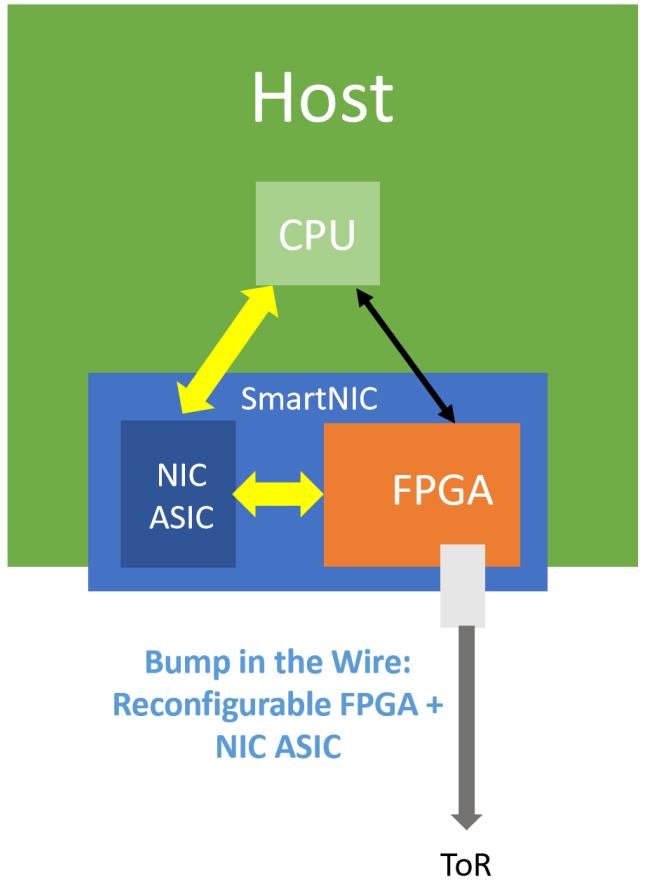


Image from D. Firestone et al., NSDI 2018

Resource Access Requirements

- Heterogeneous resources
  - CPU
  - Memory
  - Other FPGAs
  - GPU
  - SSD

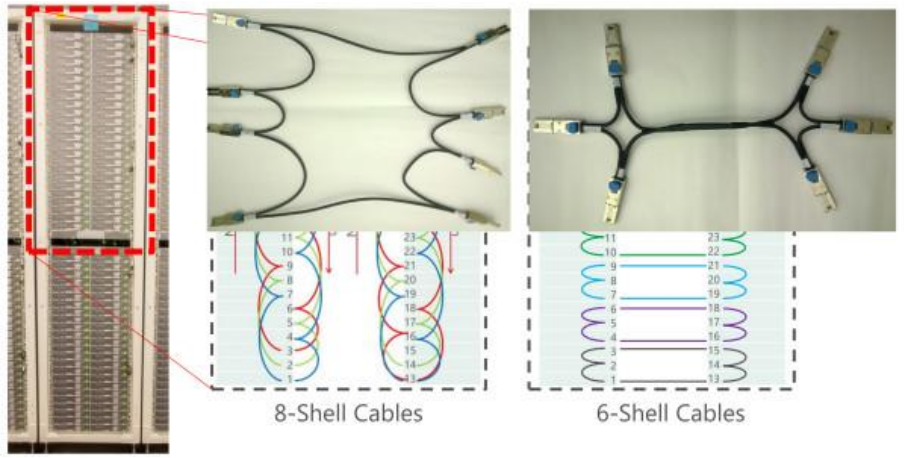
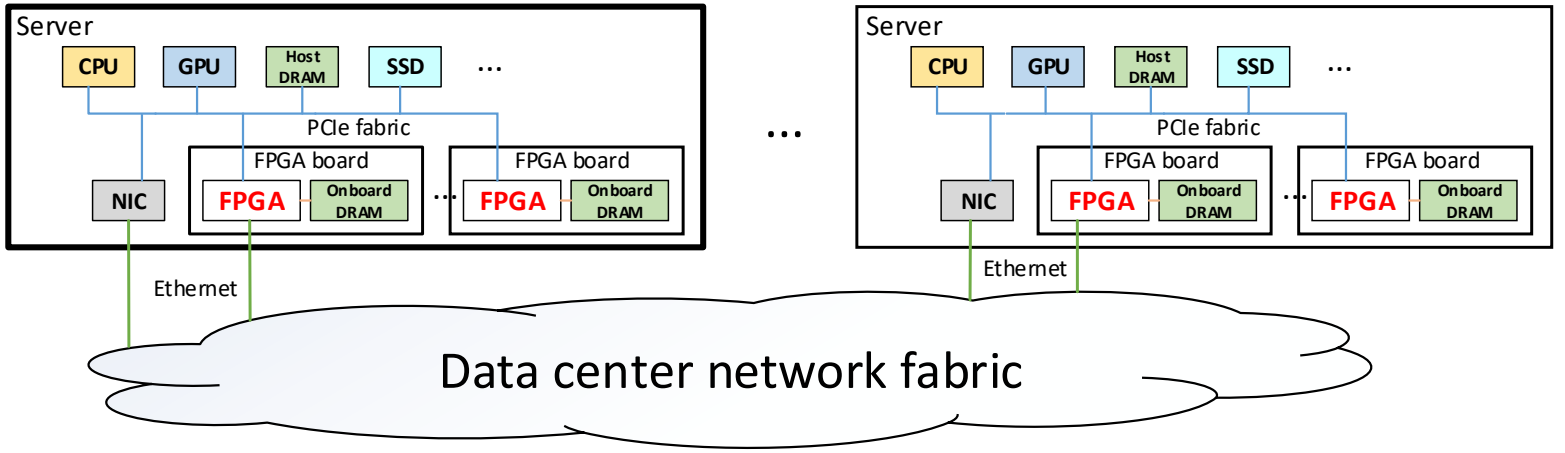


Image from A. Putnam et al., ISCA 2014 [3]



FPGA Board in Data Center

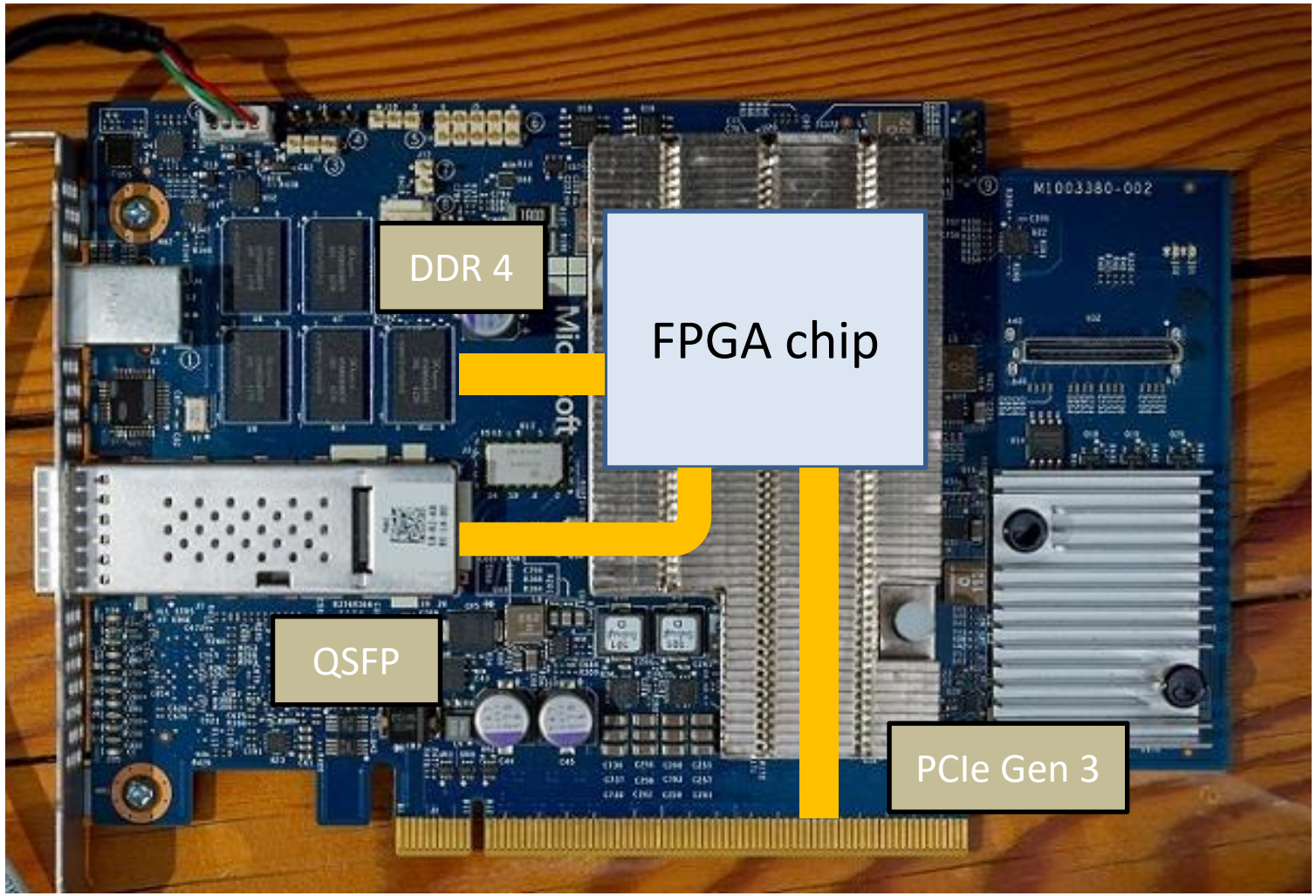
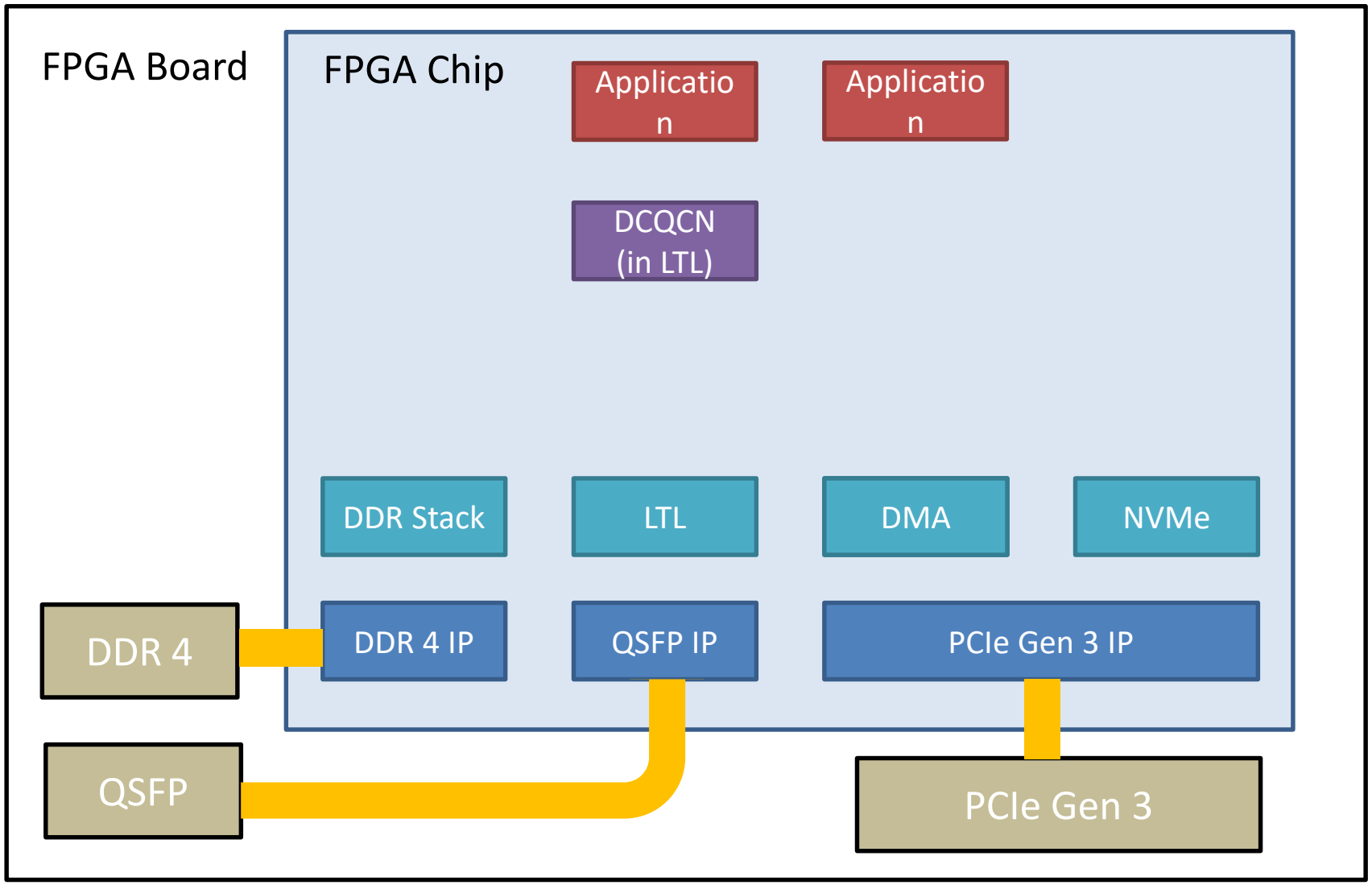


Image from https://www.cnet.com/news/microsoft-project-brainwave-speeds-ai-with-fpga-chips-on-azure-build-conference/

Current FPGA Communication Architecture



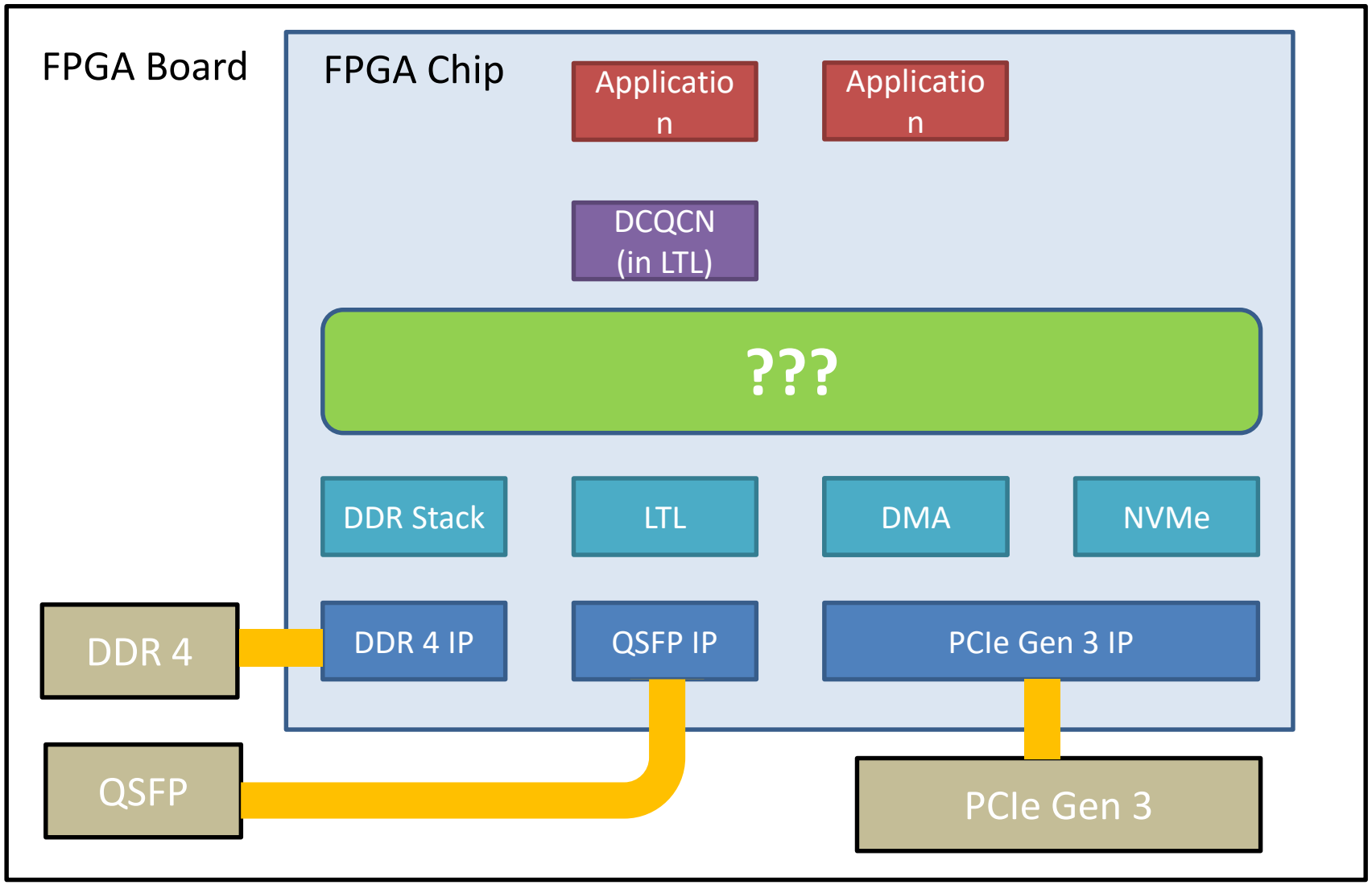
Application Layer

Transport Layer

Data Link Layer

Physical Layer

Current FPGA Communication Architecture



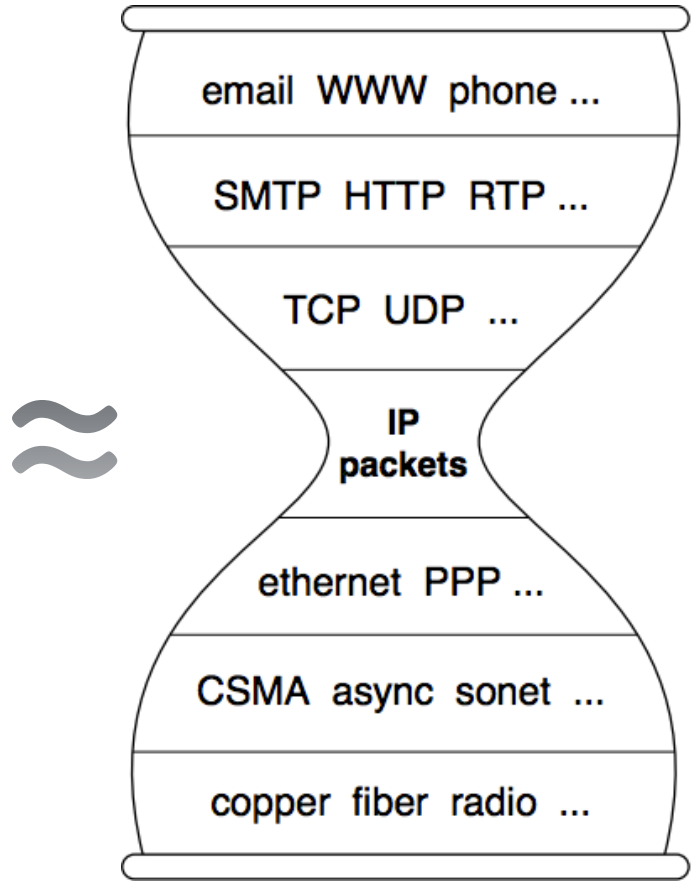


Image from L. Zhang et al., CCR 2014

Problem #1 – Programming Interface

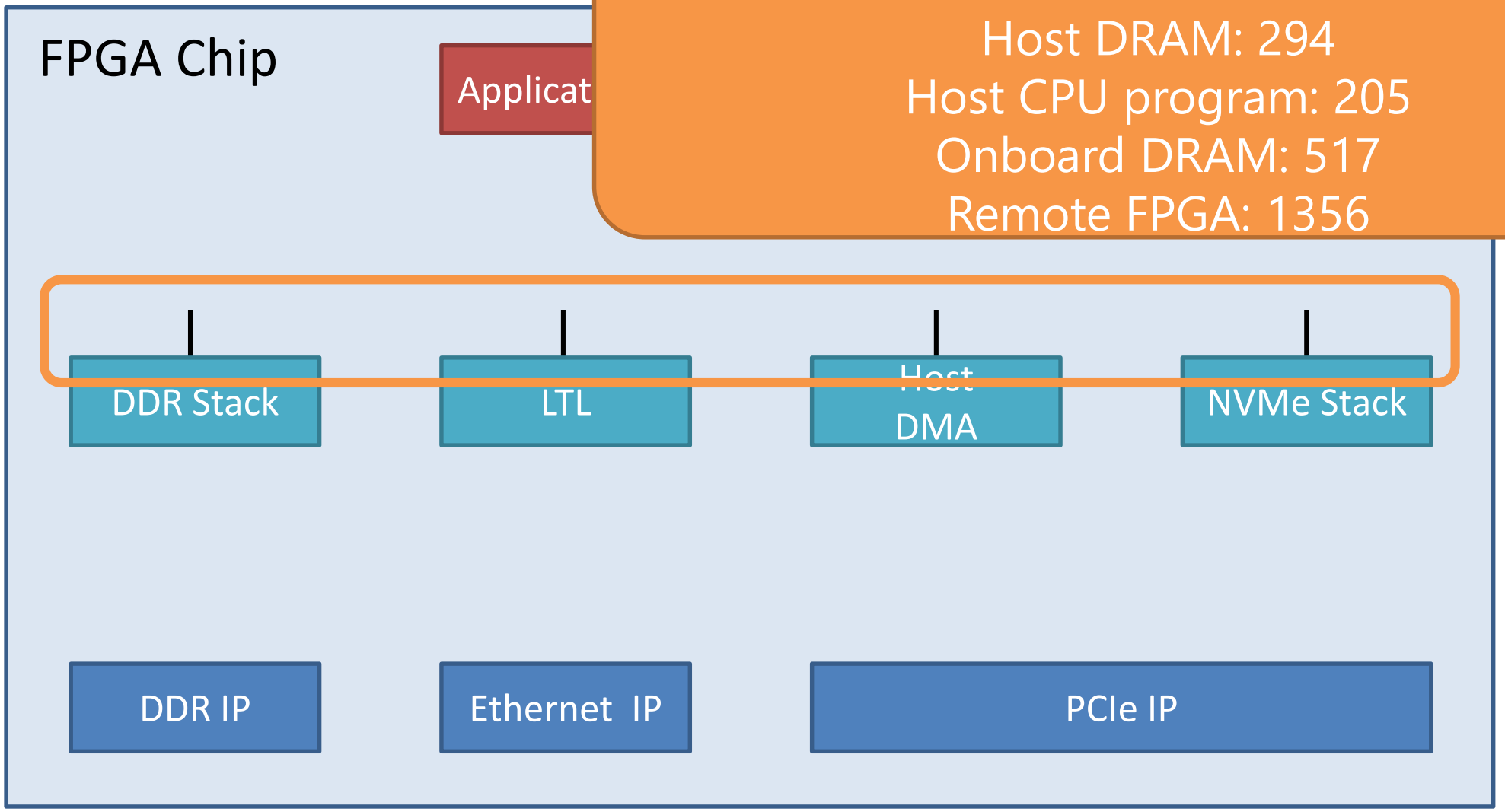
Lines of code to use each interface

Host DRAM: 294

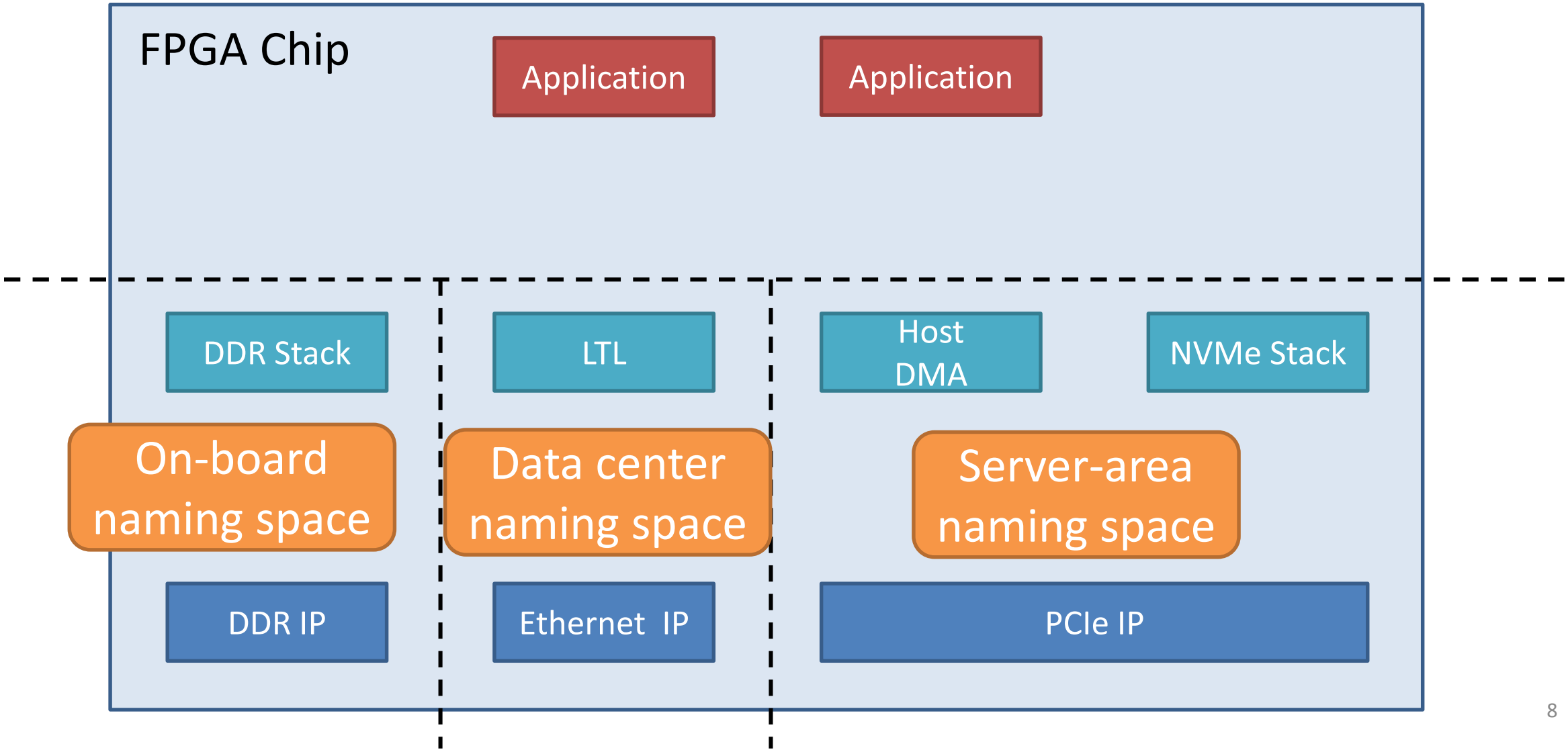
Host CPU program: 205

Onboard DRAM: 517

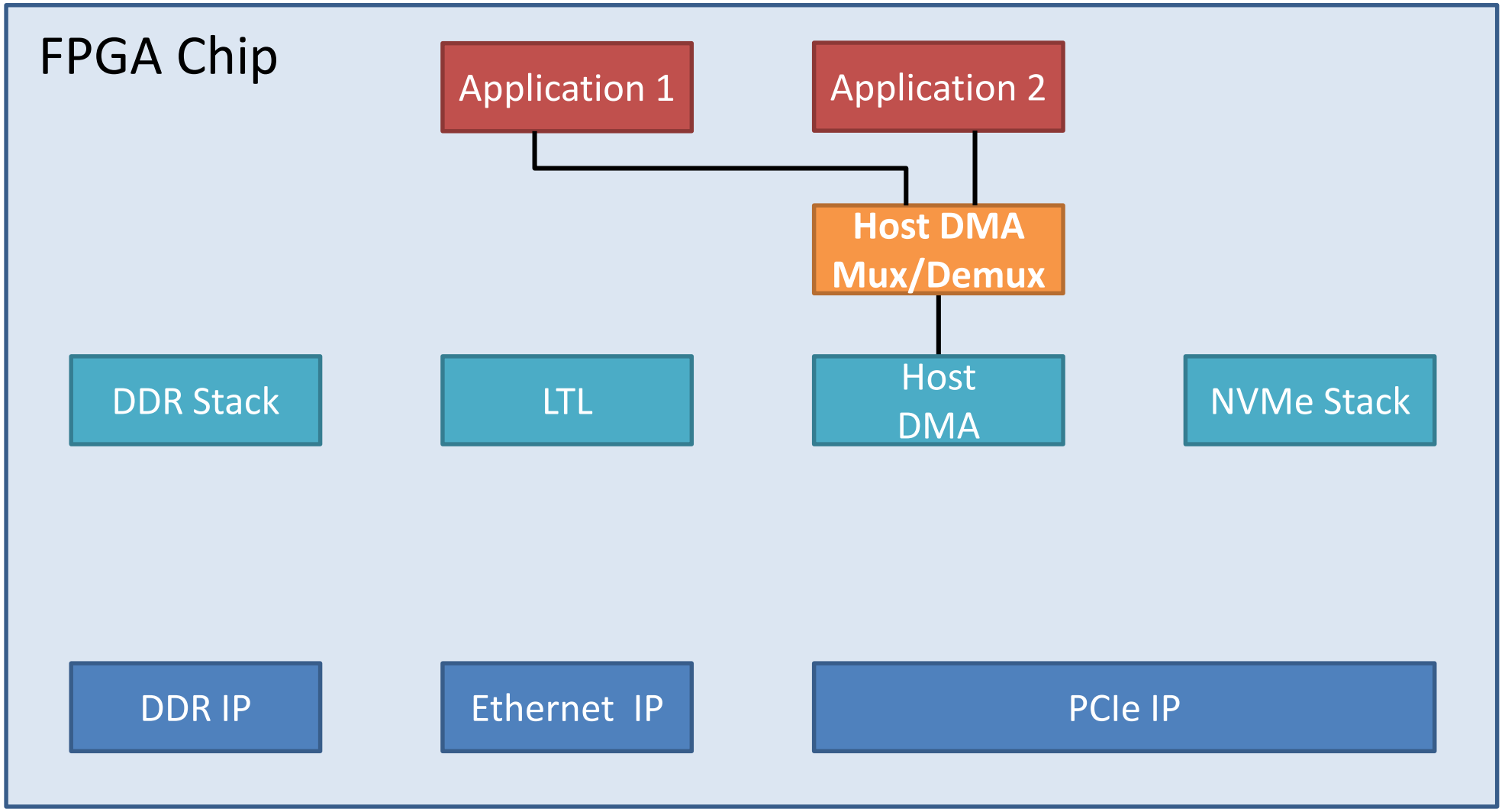
Remote FPGA: 1356



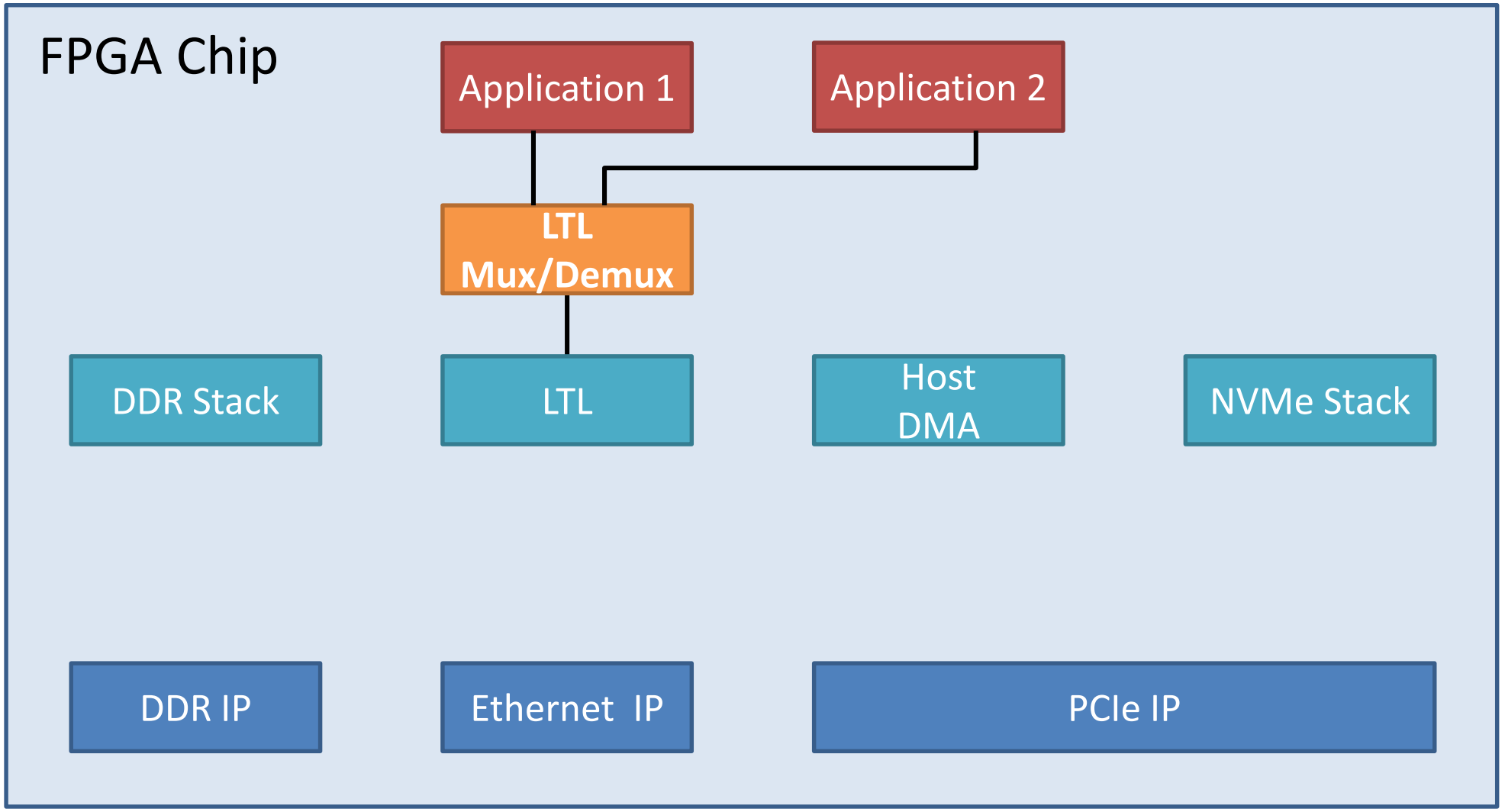
Problem #2 – Accessibility



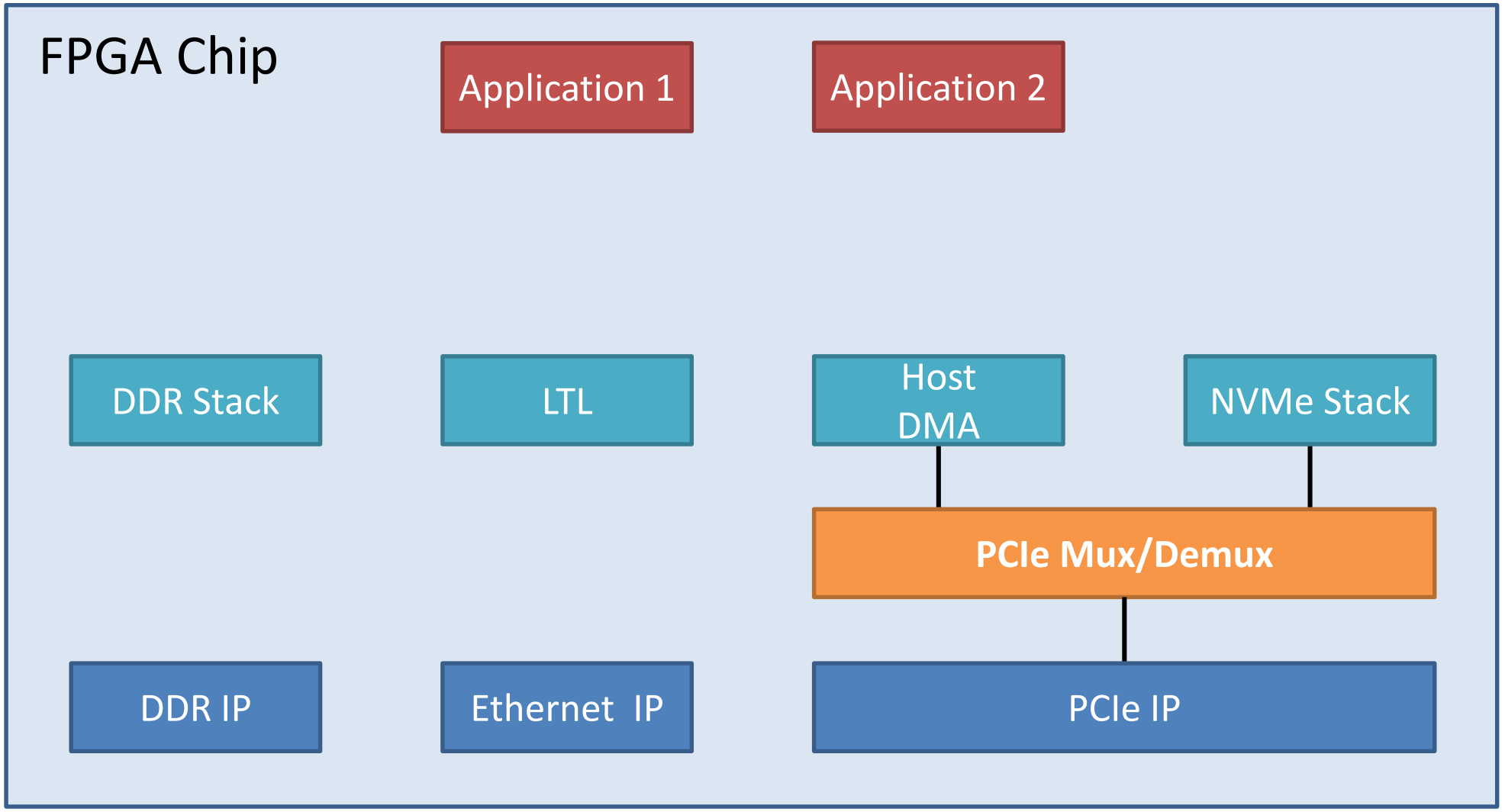
Problem #3 – Multiplexing



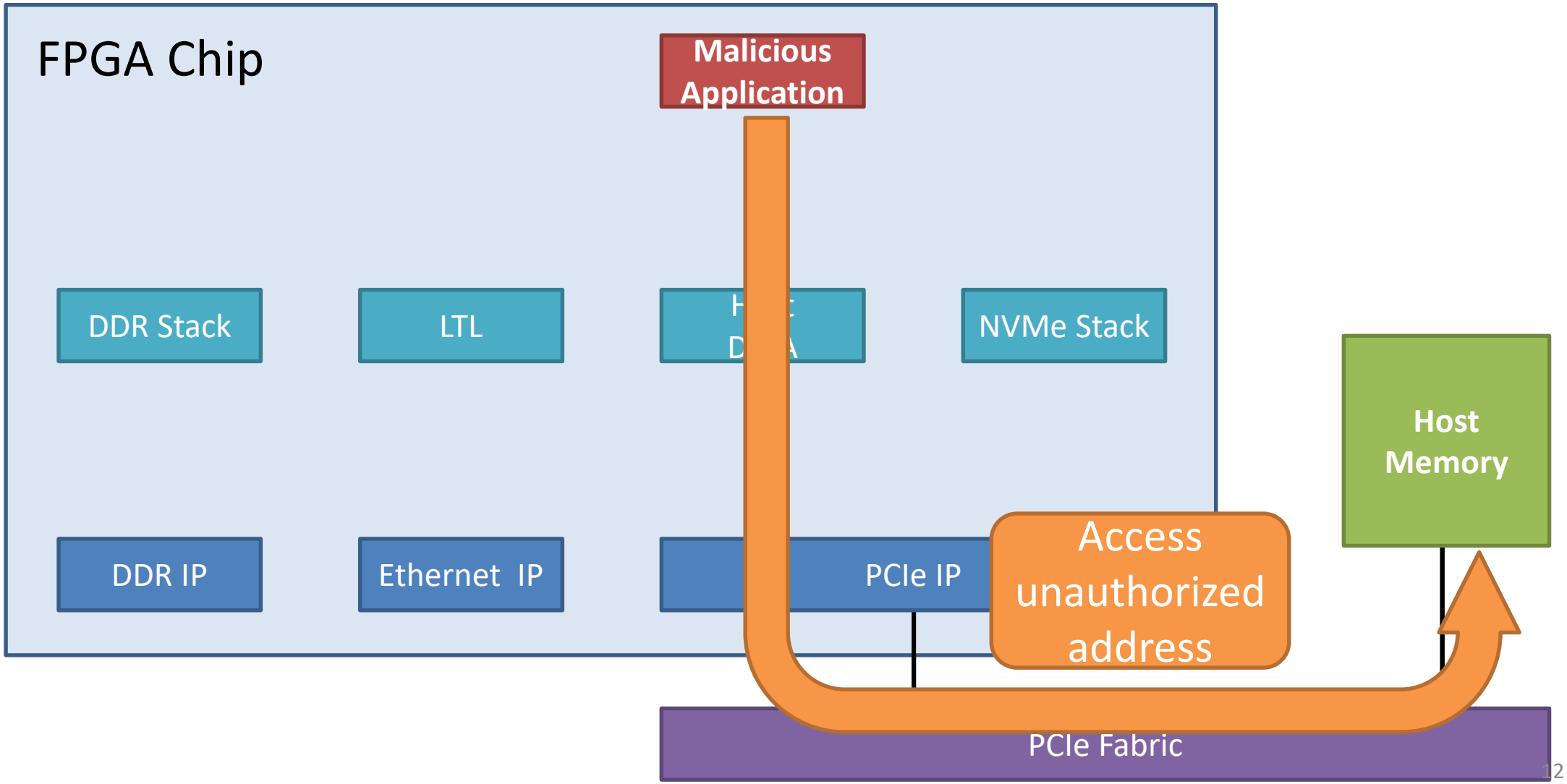
Problem #3 – Multiplexing



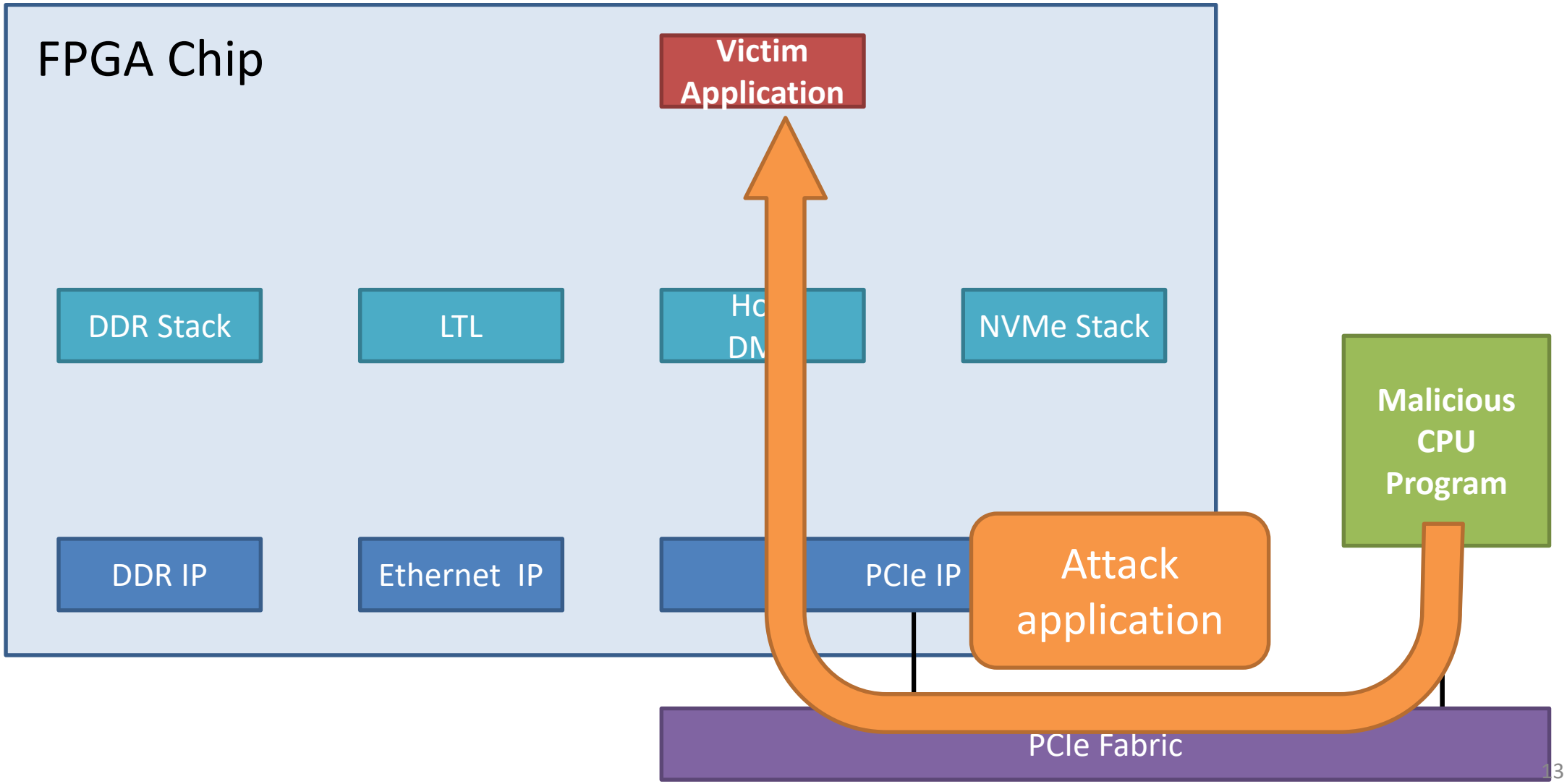
Problem #3 – Multiplexing



Problem #4 – Security



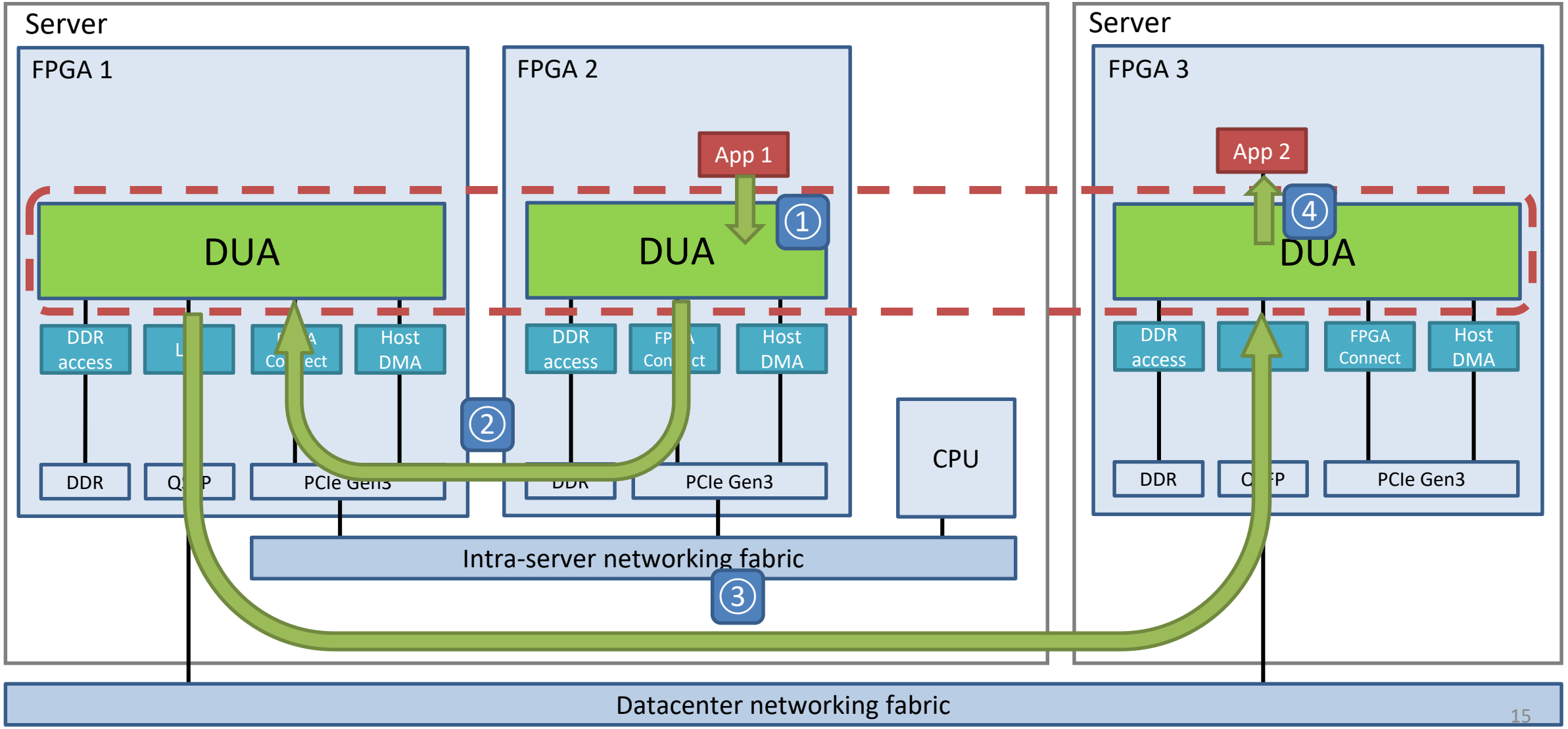
Problem #4 – Security



Existing Problems

- Complex programming interface
- Separate naming space
- No general multiplexing
- Security issue

Direct Universal Access



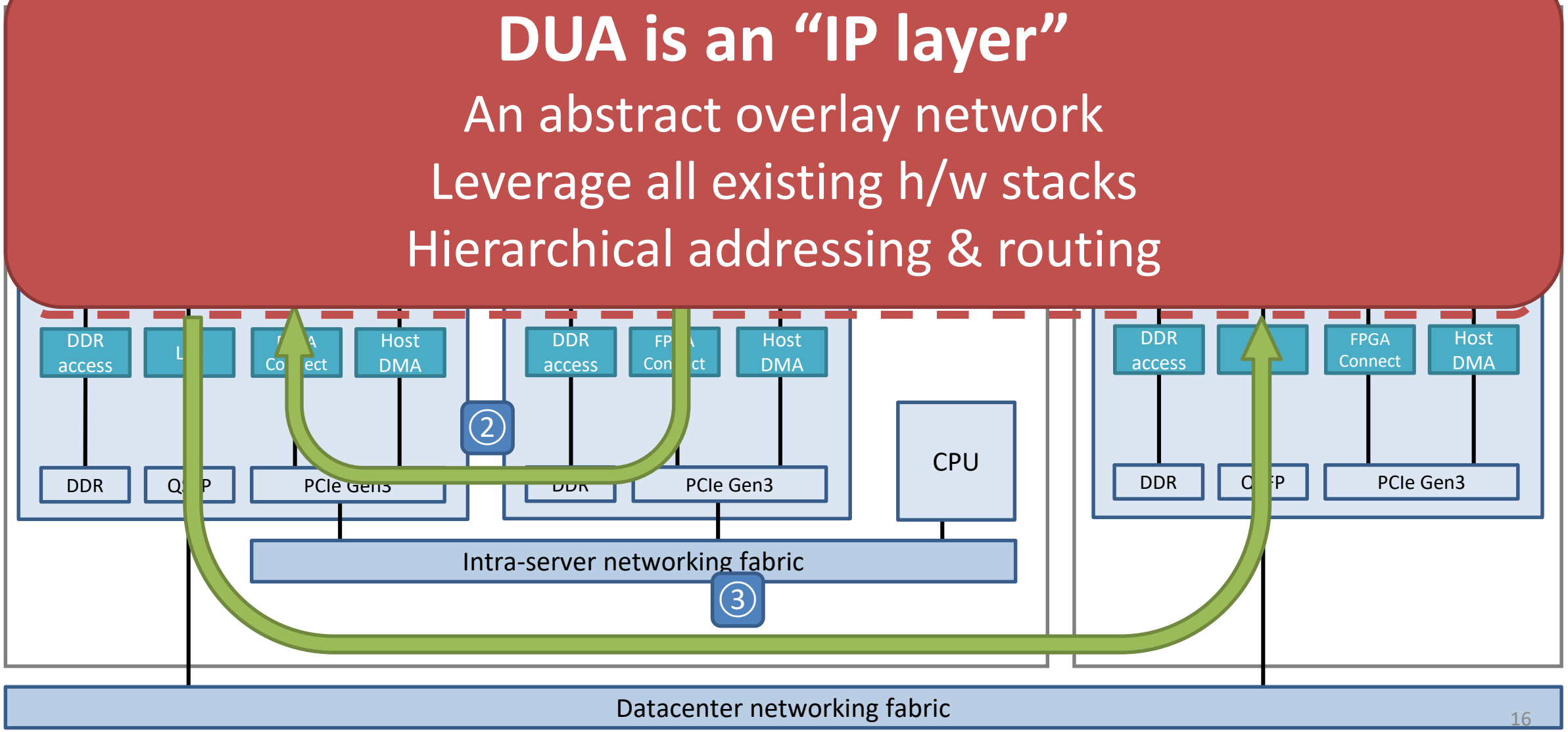
DUA Overview

DUA is an "IP layer"

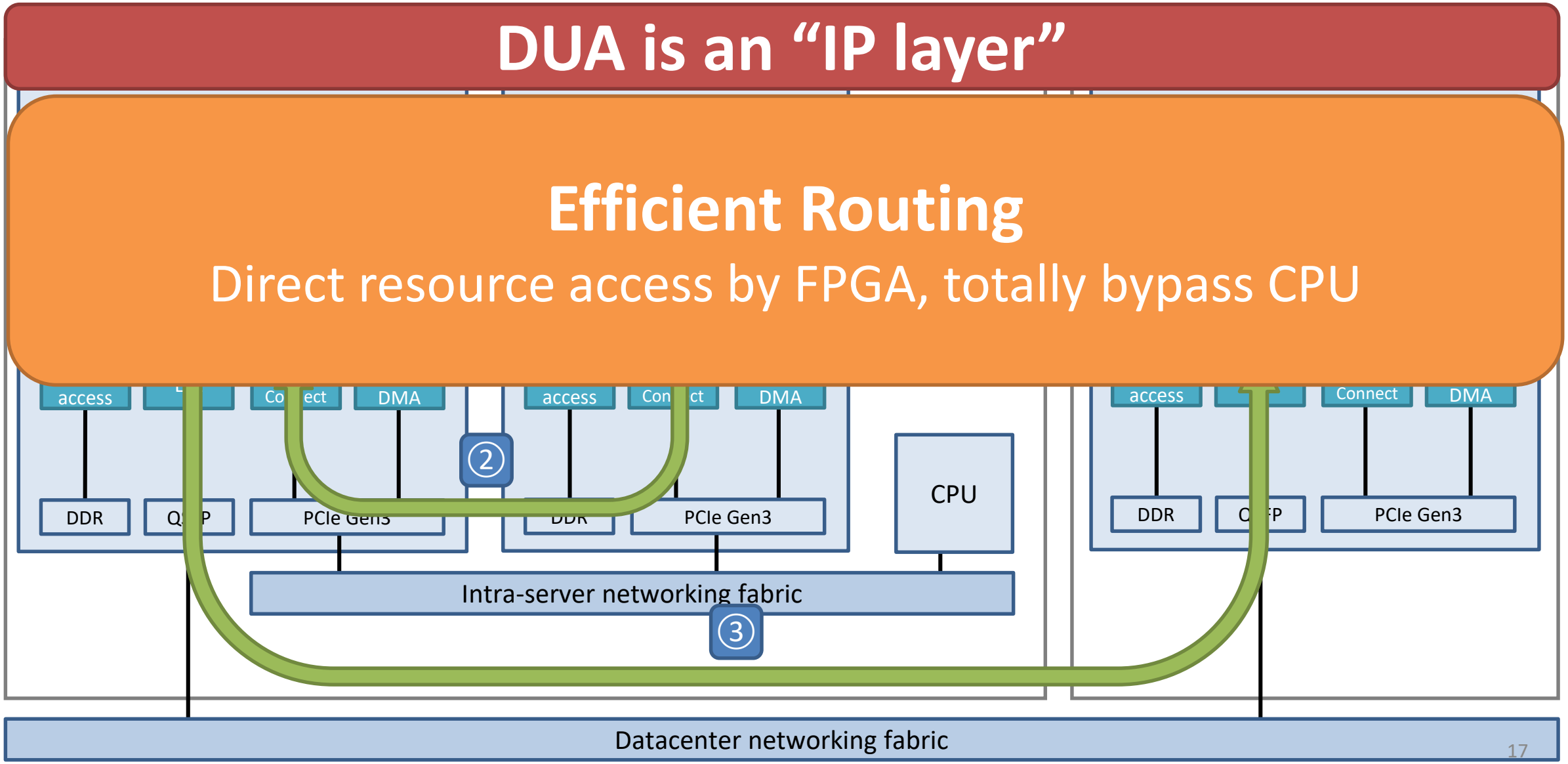
An abstract overlay network

Leverage all existing h/w stacks

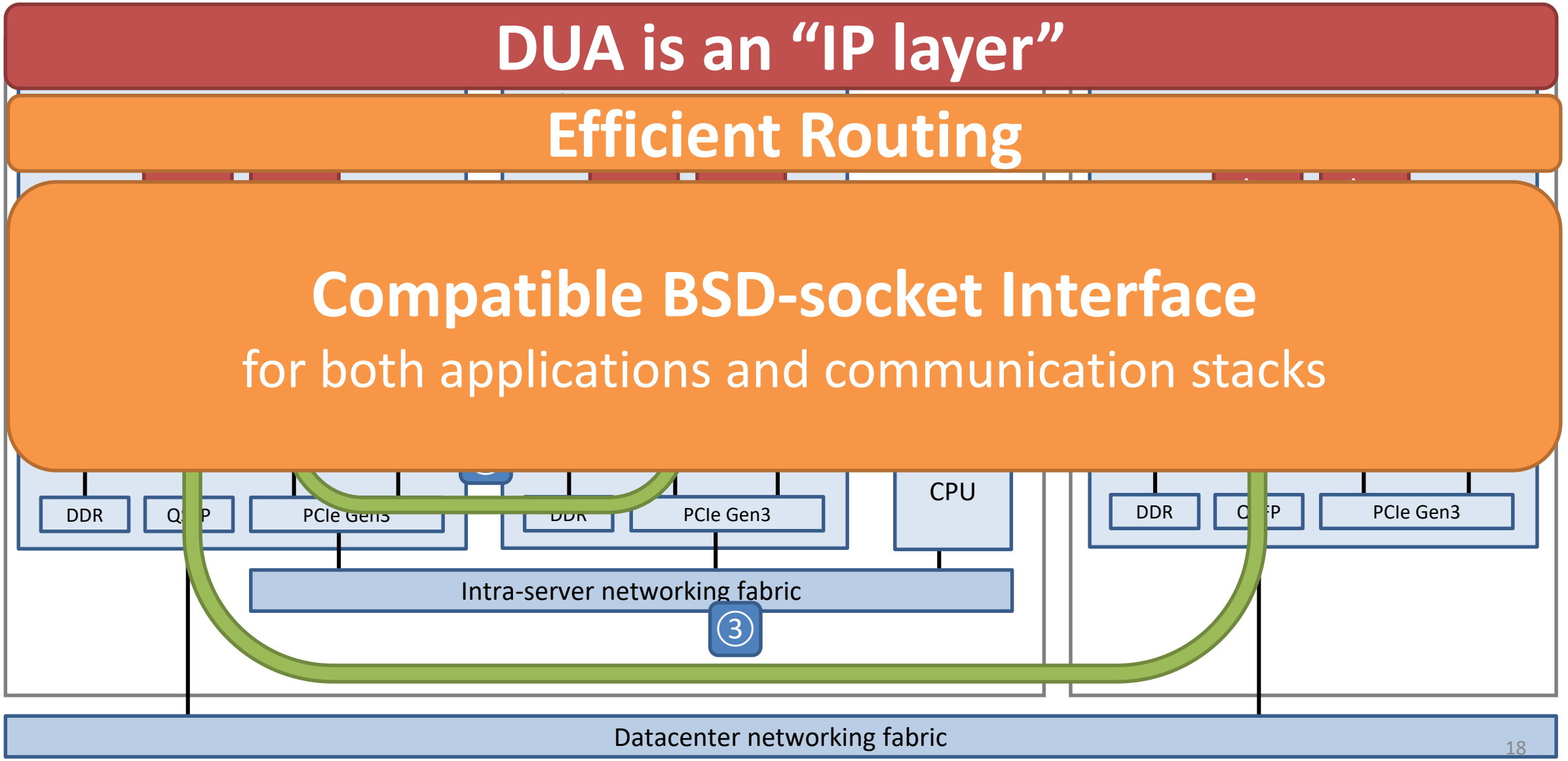
Hierarchical addressing & routing



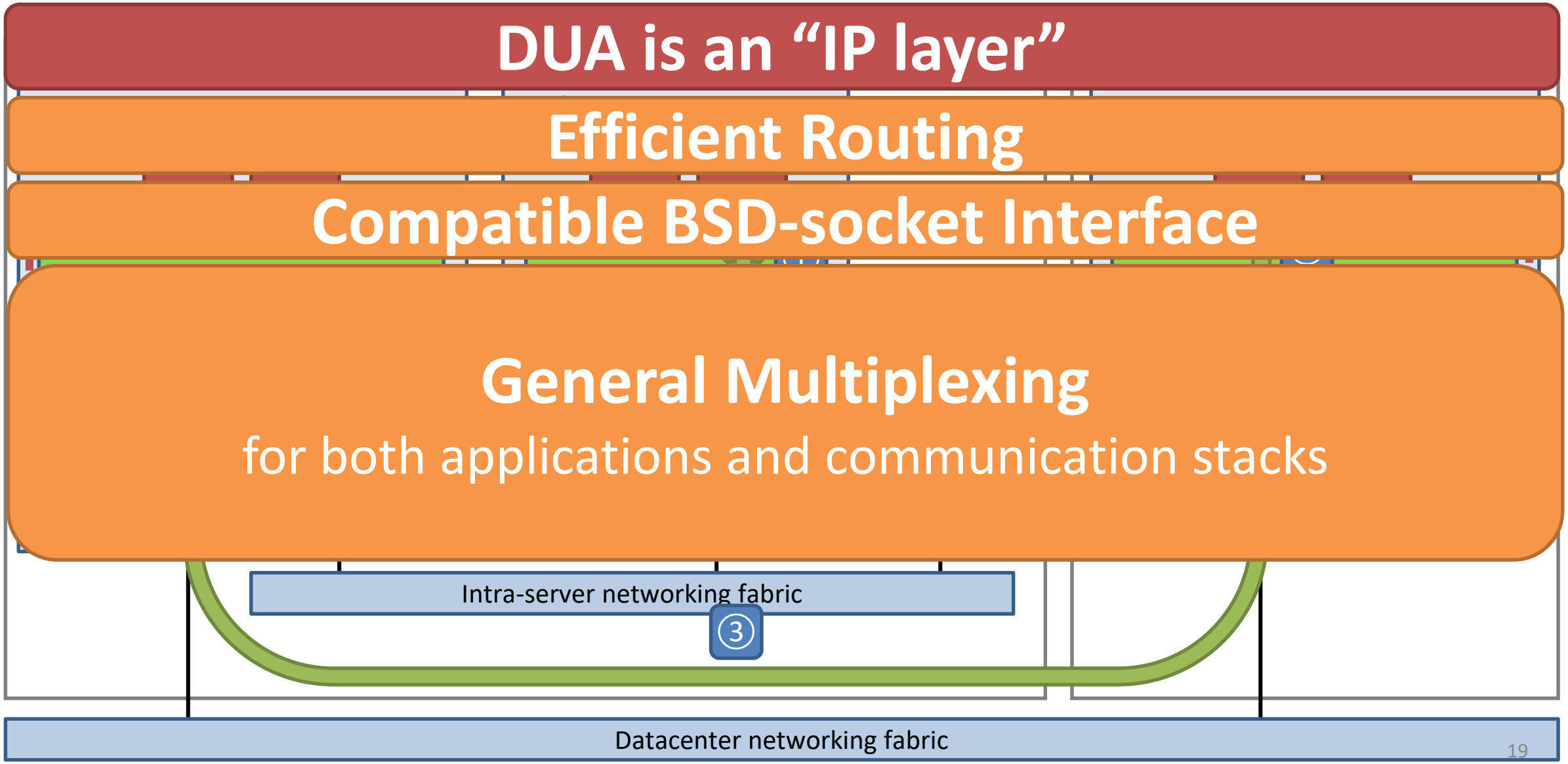
DUA Overview



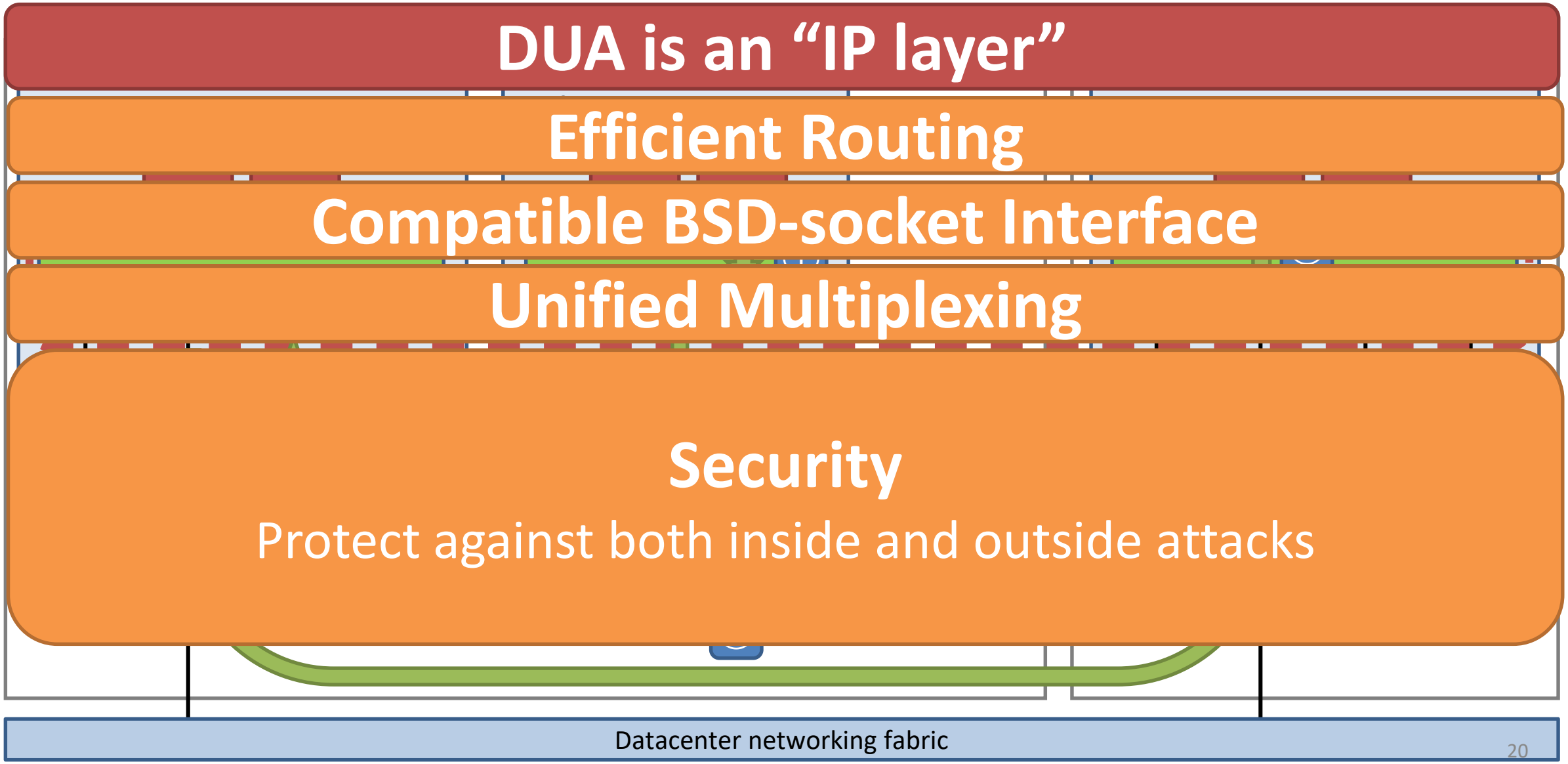
DUA Overview



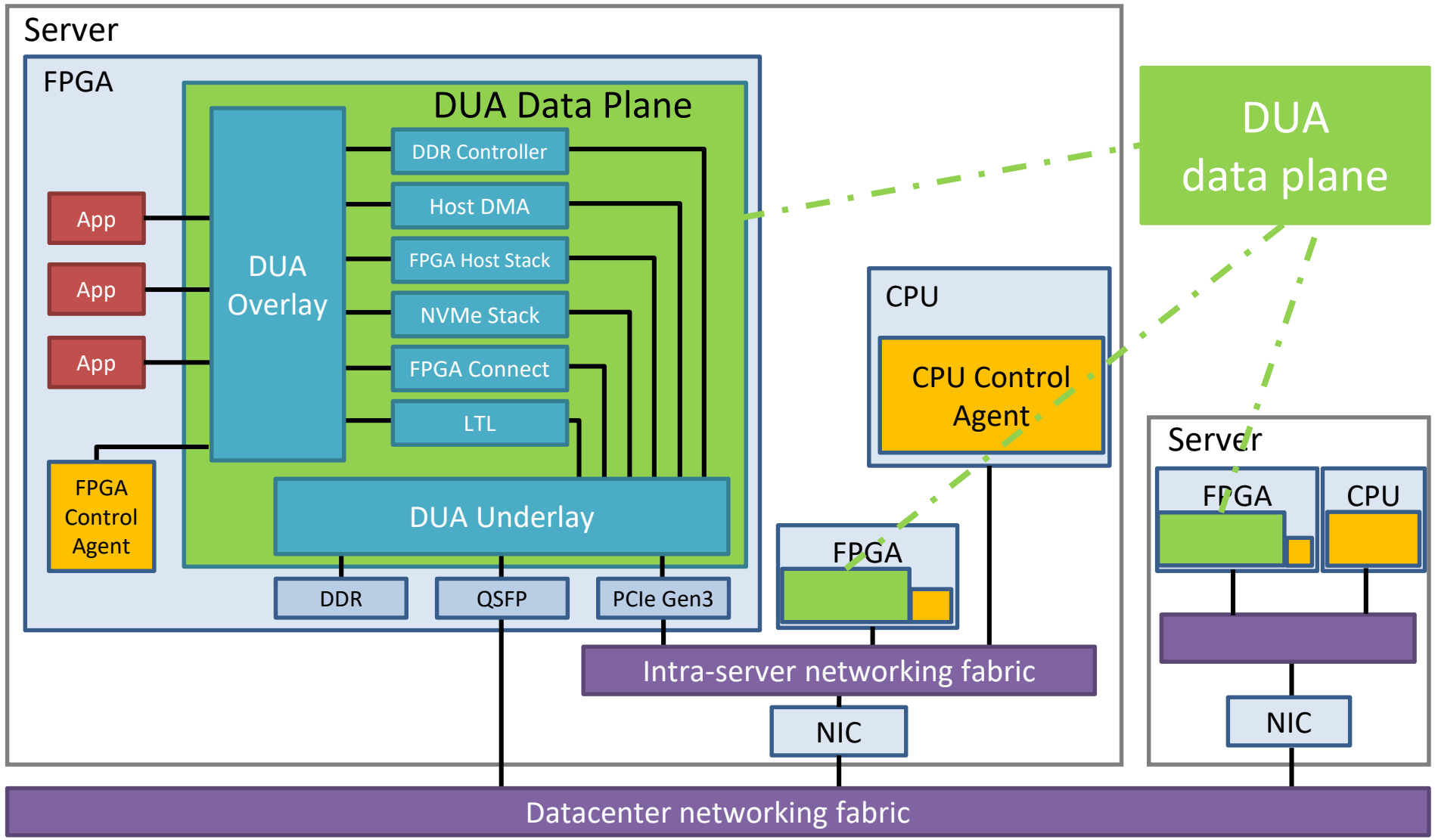
DUA Overview



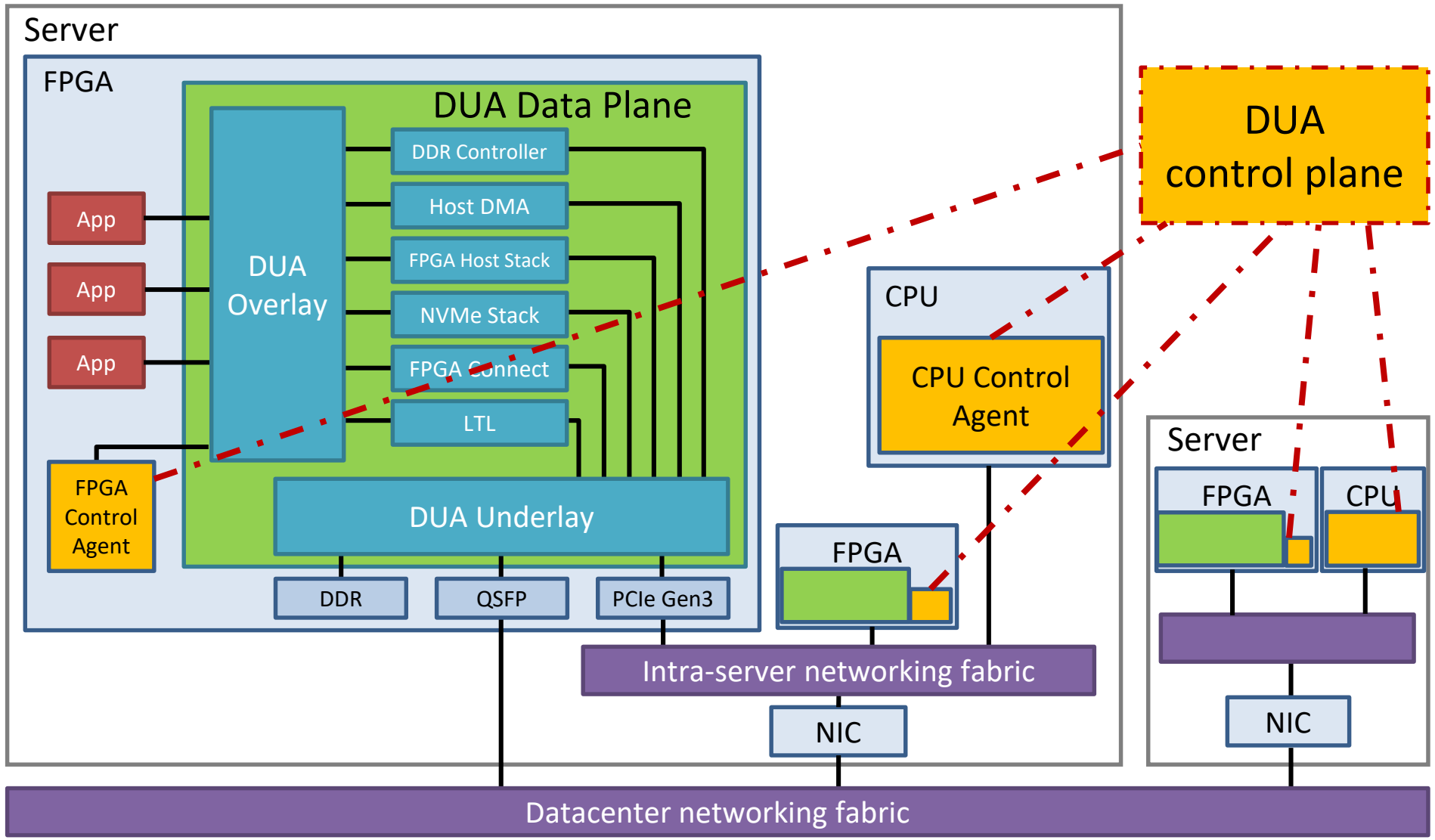
DUA Overview



System Architecture



System Architecture



DUA Control Plane

- Challenge: large-scale resource and routing info dissemination
  - Limited h/w resource
- DUA solution
  - Hierarchical addressing
  - Hierarchical routing
  - Leverage existing infrastructure
- Fully distributed and lightweight
  - **Need no global synchronization**

UID (serverID:deviceID)	Address /port	Resource description
192.168.0.2:1	0x00000001CFFFF000	1st block of host DRAM
192.168.0.2:1	0x000000019FFFF000	2nd block of host DRAM
192.168.0.2:2	0x80000000	1st block of FPGA onboard
192.168.0.2:3	8000	1st application on FPGA
192.168.0.2:3	8001	2nd application on FPGA

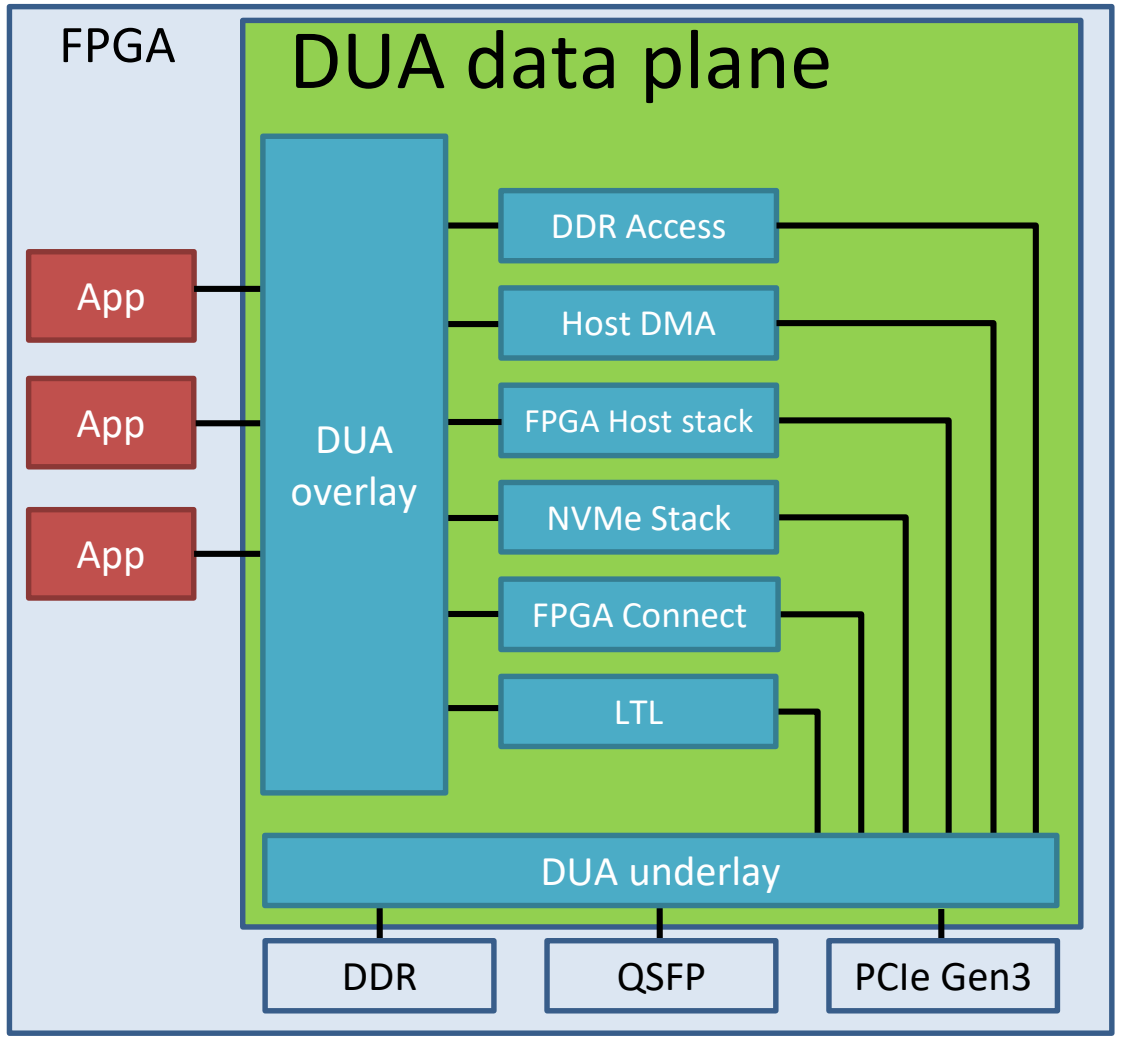
Resource table

Src Resource (UID)	Dst Resource (UID) / Stack
FPGA 1 (192.168.0.2:1)	FPGA 2 (192.168.0.2:2) / FPGA Connect
	Host DRAM (192.168.0.2:3) / DMA
	Onboard DRAM (192.168.0.2:4) / DDR
FPGA 2 (192.168.0.2:2)	FPGA 1 (192.168.0.2:1) / FPGA Connect
	Host DRAM (192.168.0.2:3) / DMA
	Resources on other servers (*:*) / LTL

Interconnection table

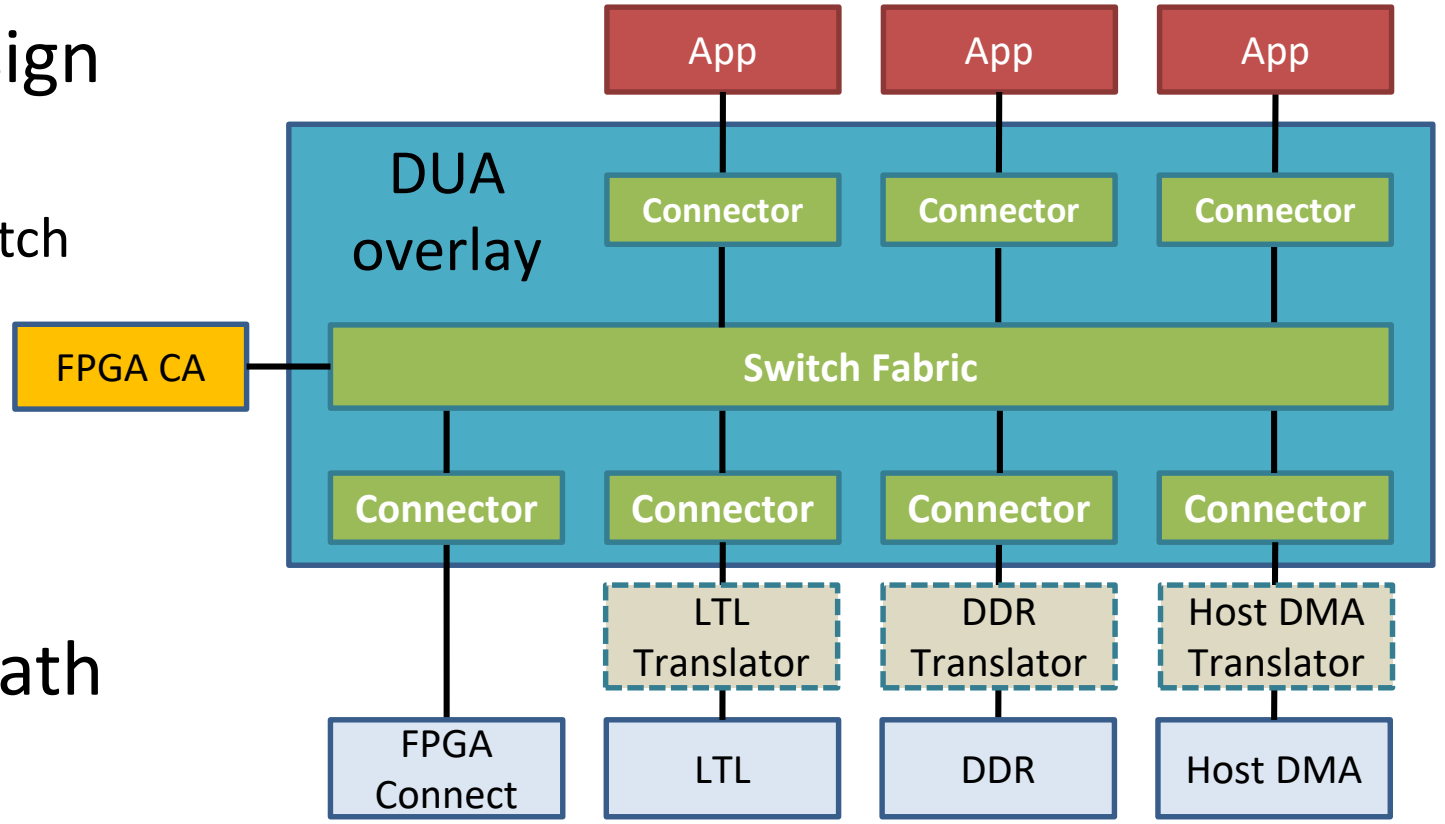
DUA Data Plane

- Overlay
  - Unified interface
  - Routing
- Stacks
  - Leverage all the existing (or adopt future) stacks
- Underlay
  - Efficient multiplexing
  - Security

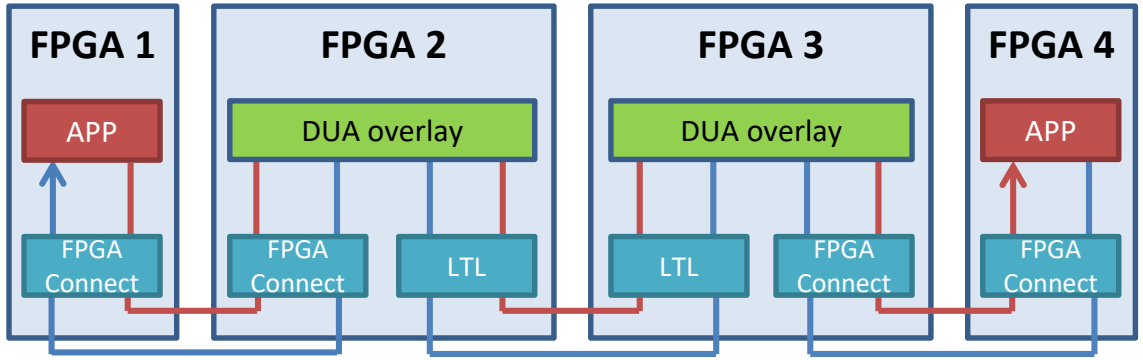


DUA Data Plane – Overlay

- Efficient & extensible design
  - Switch fabric
    - High capacity cross-bar switch
  - Connector
    - All cached routing tables
  - Translator
    - Protocol translation
- High performance data path
  - Line-rate
  - Near zero-delay

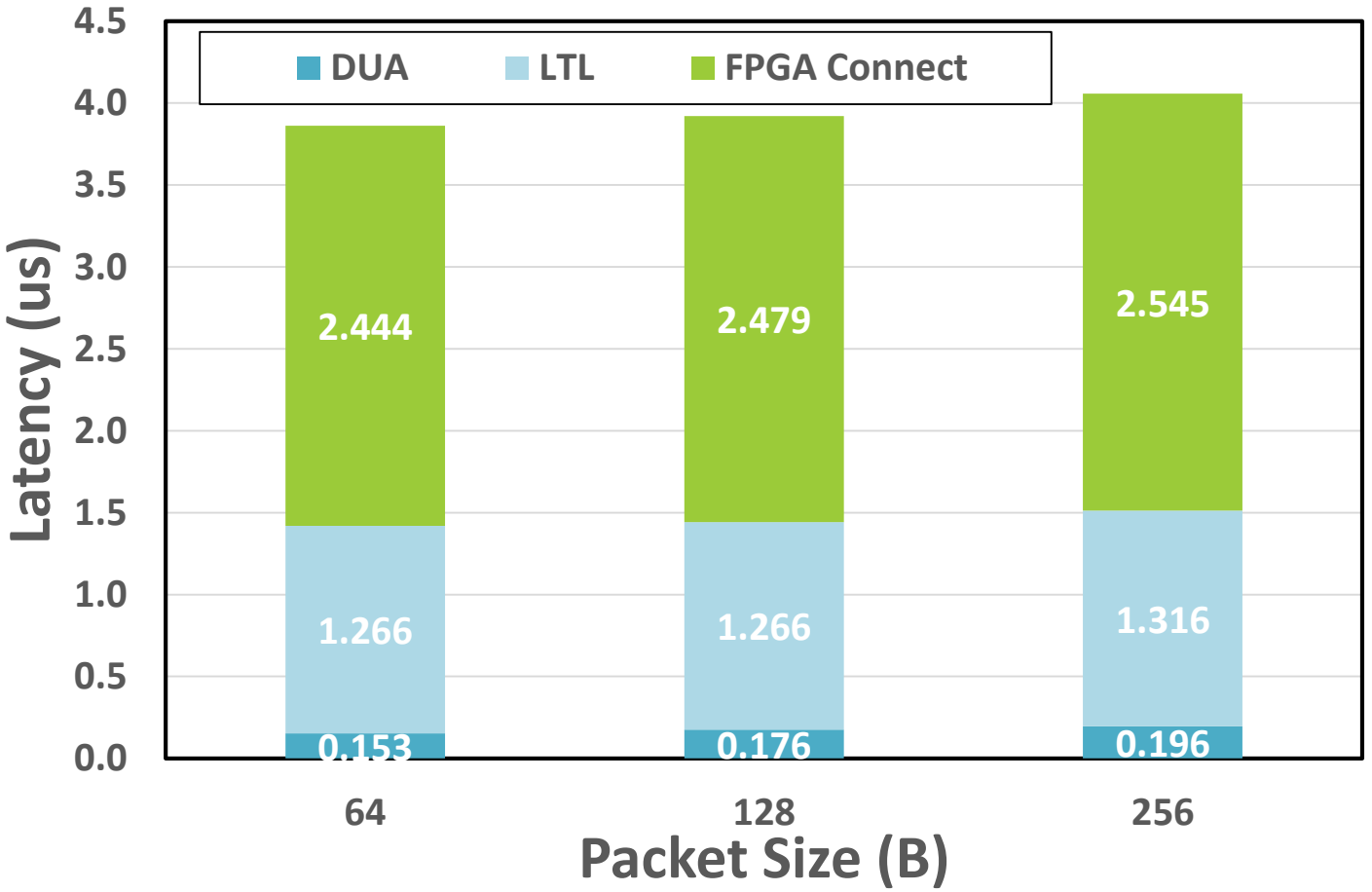


Evaluation – efficiency



Round Trip Time through FPGA Connect and DUA for 4 times, LTL twice

Extreme low latency (< 50 ns/fwd)



Evaluation – Logic Overhead

Component			ALMs	
DUA overlay	Switch fabric	2 ports	1272	0.74%
		4 ports	3227	1.88%
		8 ports	9366	5.45%
	Connector		3011	1.75%
	Stack translator	FPGA Connect	138.4	0.08%
		LTL	255.4	0.15%
		DMA	115.7	0.07%
		DDR	190.3	0.11%
DUA underlay	Stacks: FPGA Connect, LTL, DMA, DDR PHY interfaces: PCIe, DDR, QSFP		431.7	0.25%

Component		ALMs	
Stack	FPGA Connect	620.8	0.36%
	LTL	6395.4	3.72%
	DMA	1347.7	0.78%
	DDR	73.4	0.04%
PHY interfaces	PCIe	3890.1	2.26%
	QSFP	12726.7	7.40%
	DDR	7369.2	4.28%

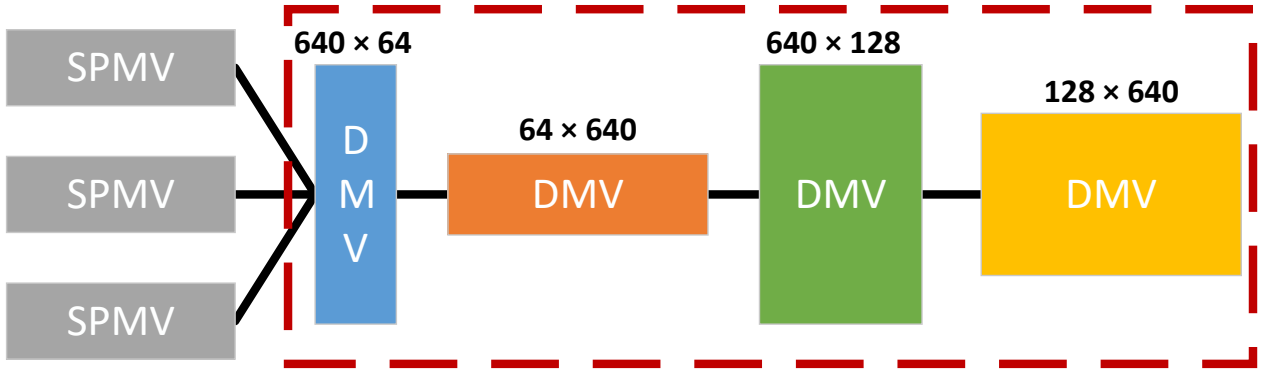
DUA Overlay

- 2 Ports: **4.24%**
- 4 Ports: **9.29%**
- 8 Ports: **19.86%**

DUA Underlay

- 4 Stacks and 3 PHY Interfaces: **0.25%**

Evaluation – Deep Crossing

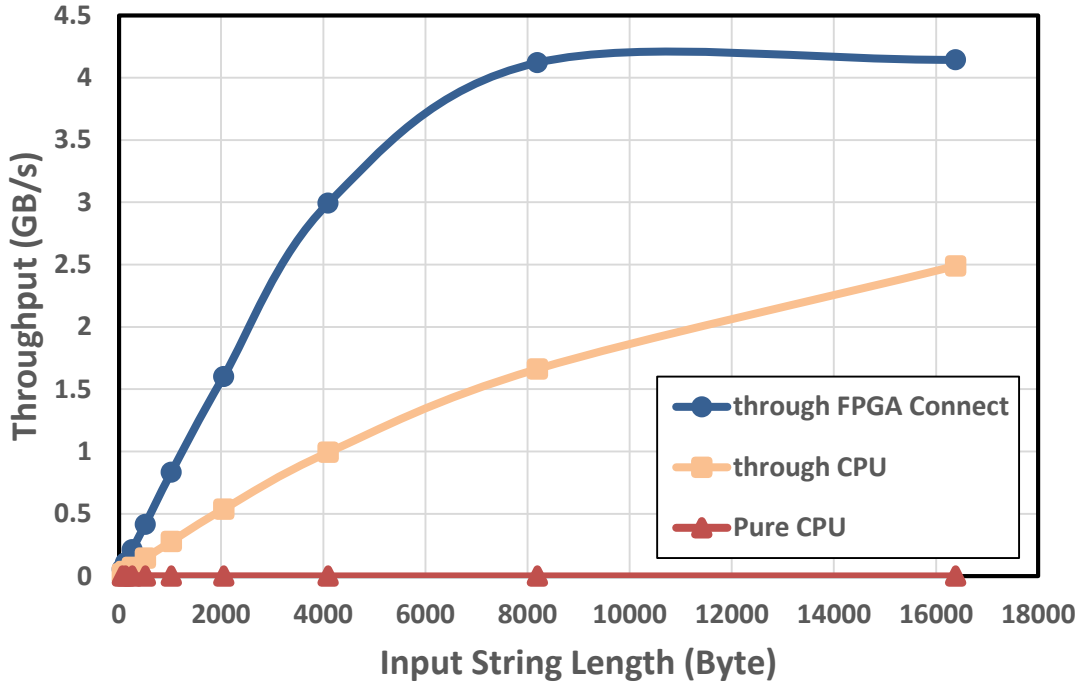


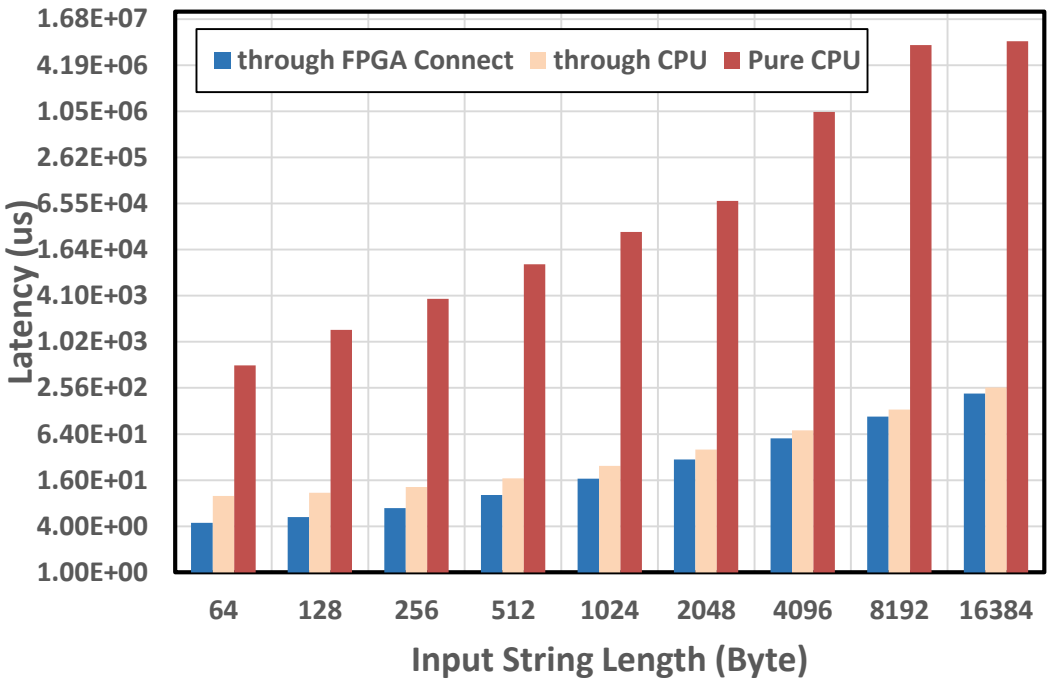
	D1	D2	D3	D4	All
Parall = 32	7.27	6.58	13.30	12.96	40.12
Parall = 64	4.17	3.48	7.22	7.09	**21.95**
Reduction(%)	42.67	47.10	45.75	45.33	45.28

Single FPGA Board: Parall = 32, 2 FPGA Board: Parall = 64

45.28% Latency Reduction

Evaluation – Regex Matching





- **Up to 10^5~10^7** higher than CPU, **Up to 10^5** lower than CPU
- **Up to 3 times** throughput and **up to 55%** latency reduction compared to using CPU to move data between FPGAs

Conclusion

- Current FPGA communication architecture
  - No universal access
- DUA: build the “IP” layer for FPGA in data center
  - Leverage existing data center network
  - Efficient routing
  - Compatible BSD socket interface
  - Unified multiplexing
  - Security
- Open source soon

Thank you!
Questions?