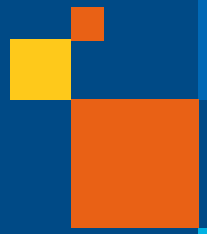
IPU/SmartNIC Accelerate Applications on the Wire

Tom Spencer
Business development
Intel Corporation
October 19, 2023



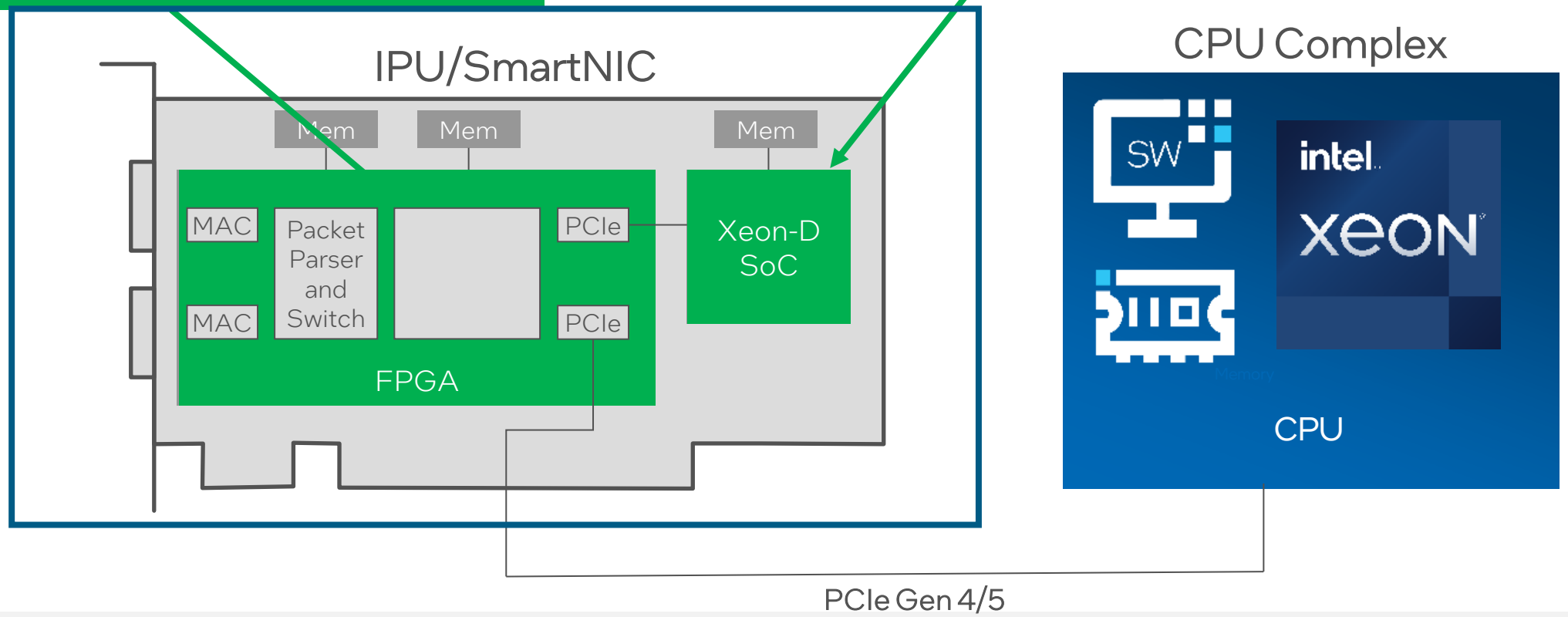
intel®

IPU/SmartNIC – Accelerate Applications on the Wire

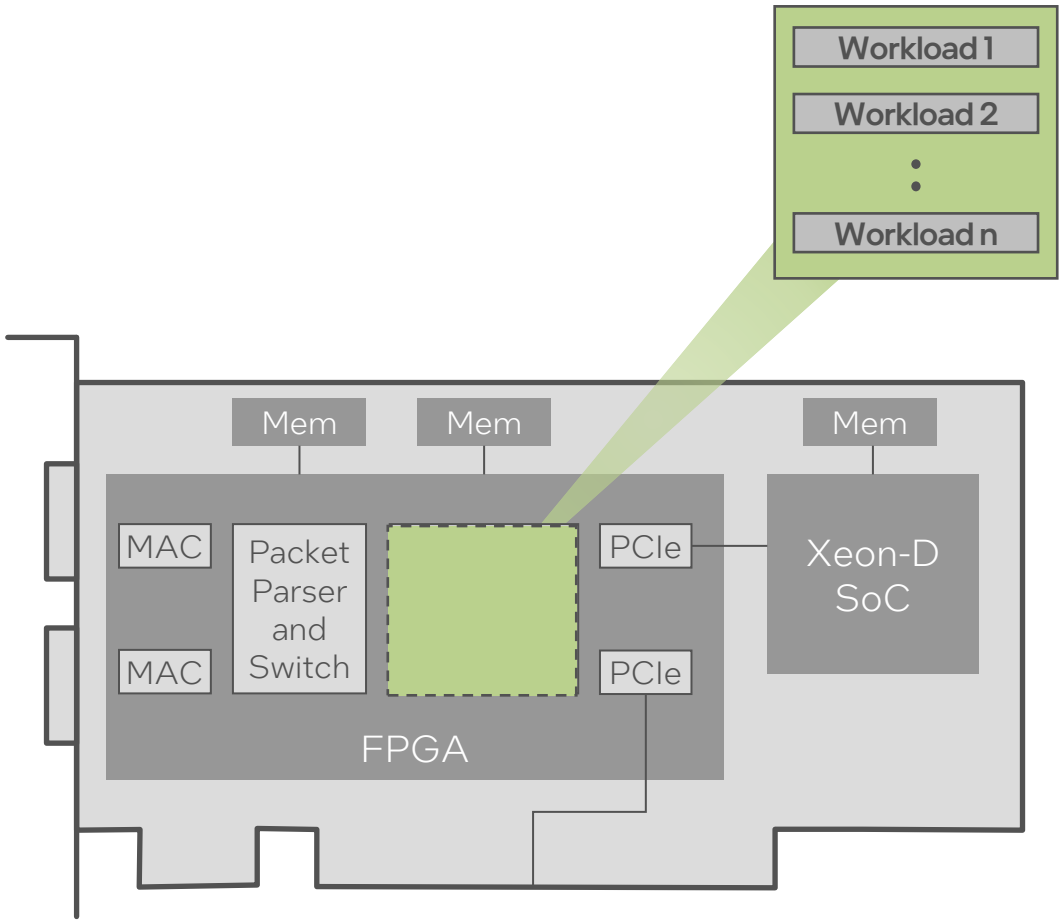
HW and SW Programmable
- Complete network stack processing
- 5 tuple parser
- Packet switch
- Optional sub-1us latency
- Embedded compute acceleration

x86 Processor
- Local OS
- Eases app migration from host

Acceleration on the Wire



Acceleration Workloads



Protocols
- TCP/IP, RoCEv2, VxLAN, Geneve, NVGRE

Network Switching/Routing
- Data plane: OvS, P4
- Analytics: Network Analytics (AI)

Security
- Encryption: TLS, KTLS, IPSec, MACSec
- Firewall → ACLs
- DDoS Mitigation

Storage
- NVMe Tunnel, NVMe-oF TCP, NVMe-oF RoCEv2, Erasure Coding, Compression

AI/ML
- xPU Interconnect
- ML/DL → LSTM, LLM

Big Data
- Arrow/Arrow Flight

Application Workloads Examples

Myrtle.ai: LSTM – Extreme Performance

- Time series analytics
- No host intervention
- VOLLO includes LSTM and Attention layers
- Proven STAC-ML benchmark data
- 48 parallel models in 1U server
- Inference leverages bfloat16 format
- PyTorch, TensorFlow, ONNX

SigmaX.ai – Big Data

APACHE ARROW + Pulsar

- Pre-processing for AI/ML
- No host intervention
- Inline Transforms/functions
- Ingest, reformat, classify (**Arrow**) and transport (**Arrow Flight**)
- Significant latency reduction

Thank You