Using FPGAs to accelerate NVMe-oF based Storage Networks

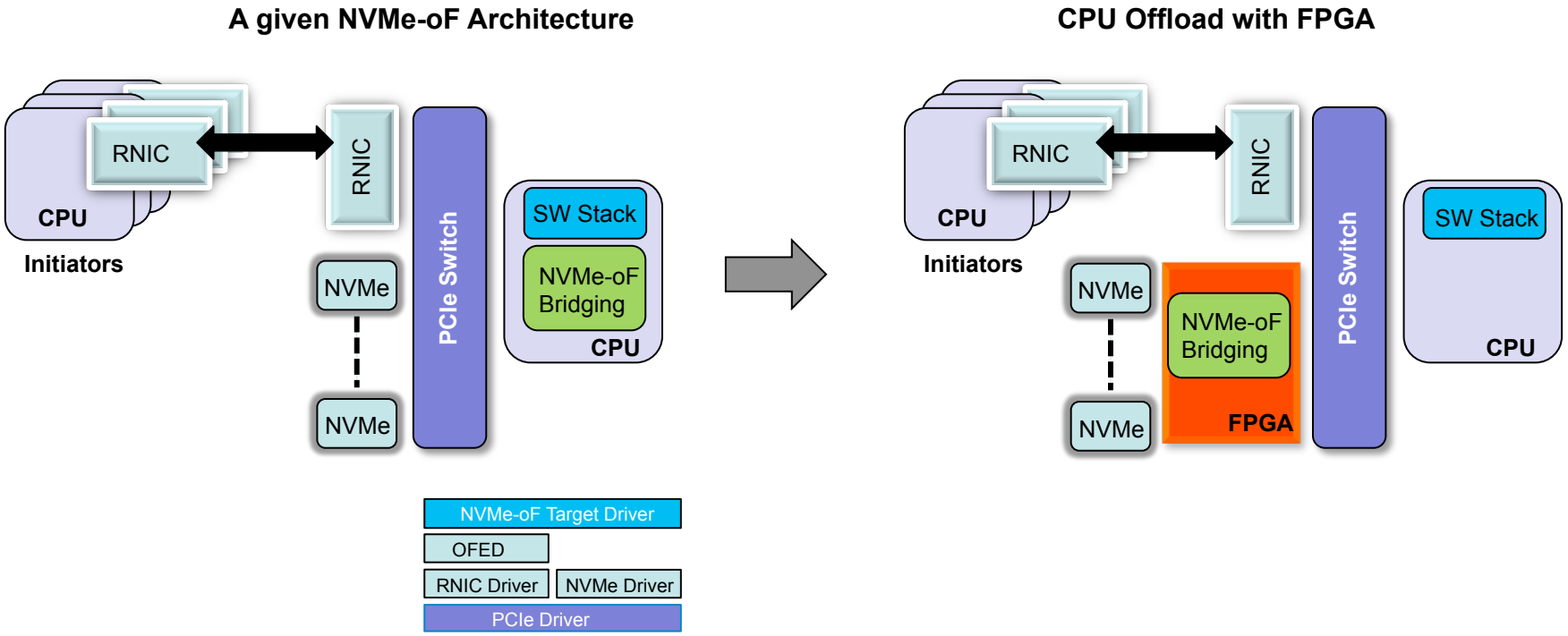
Deboleena Sakalley

IP & Solutions Architect, Xilinx

Agenda

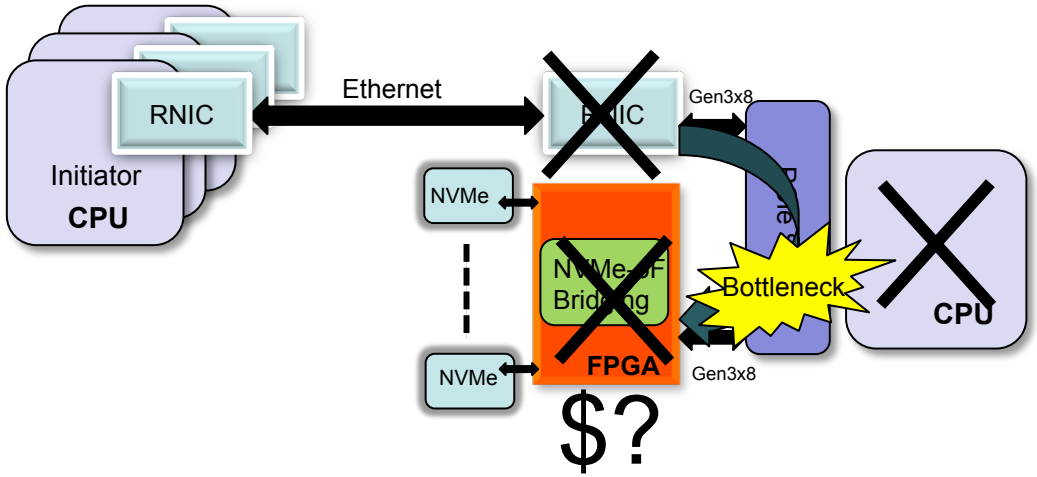
- NVMe-oF CPU Offload in FPGA
- NVMe-oF Integrated Solution
- Solution Architecture
- Performance
- Conclusions

NVMe-oF CPU Offload in FPGA



NVMe-oF CPU Offload in FPGA

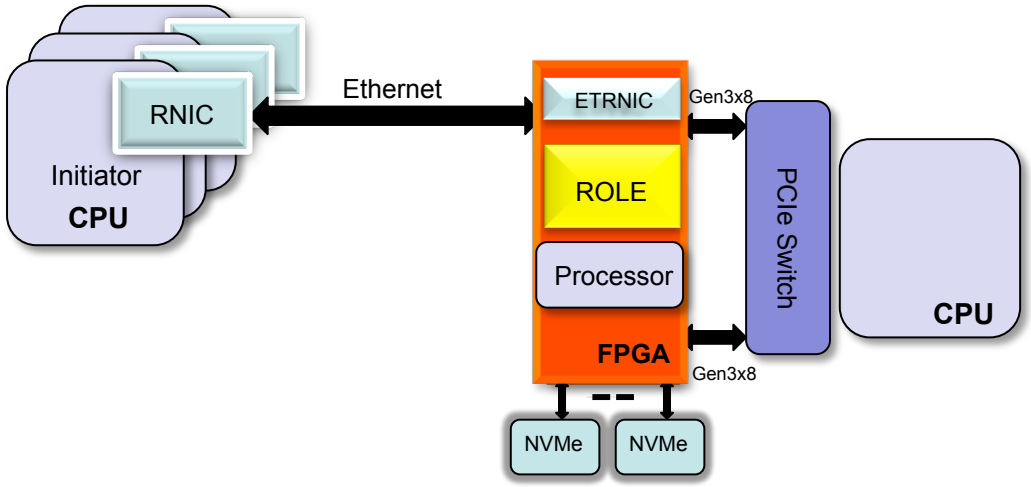
- HW accelerated NVMe-oF
- SSD sharing across multiple hosts
- CPU only in control path
  - Doorbell exchanges to/from RNIC
  - Discovery
  - Connect
  - Error handling
- Peer 2 peer transfer between RNIC and FPGA



- ❑ **P2P Bottleneck**
- ❑ **Too many components!**
- ❑ **FPGA Value Add not clear**

NVMe-oF Integrated Solution

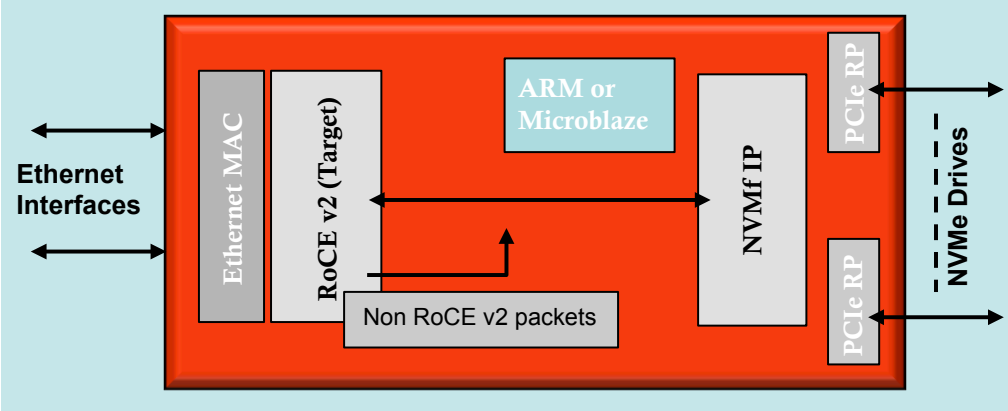
- Integrate RNIC functionality
- Use embedded FPGA processor
- Enable end-customer specific acceleration
  - RAID
  - Erasure coding
  - De-dup
  - Others



- **P2P Bottleneck**
- **Too many components!**
- **FPGA Value Add not clear**

Solution Architecture

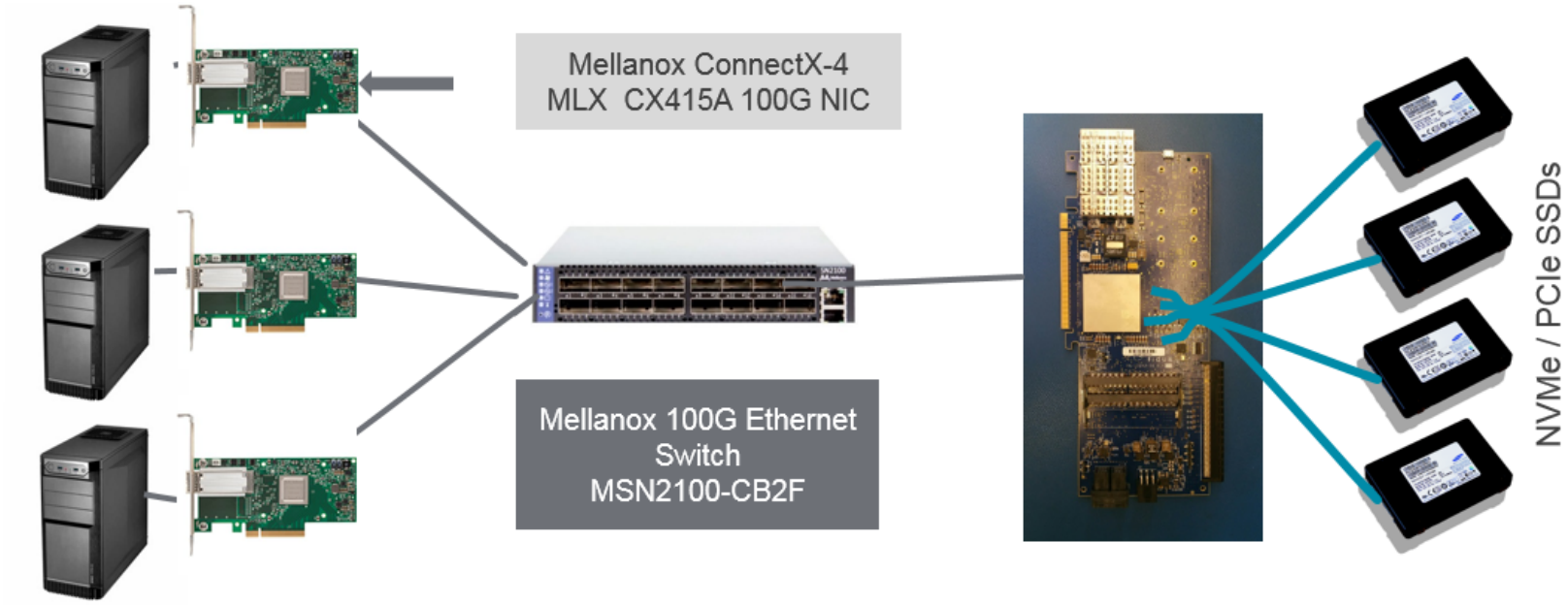
- HW implementation for Target/end point RoCE v2
- Supports up to128 initiators
- 8 to 24 flash drives (with PCIe switch)
- Support for 10/25/40/100GbE and PCIe Gen3/Gen4
- Datapath and control path handled in HW
  - All recoverable errors handled in HW
- Embedded processor only used for
  - Discovery
  - Connect
  - Error handling



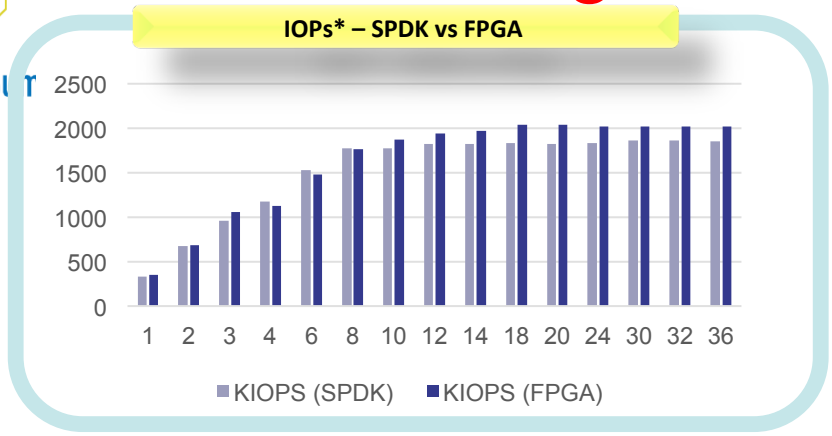
Target SW Components

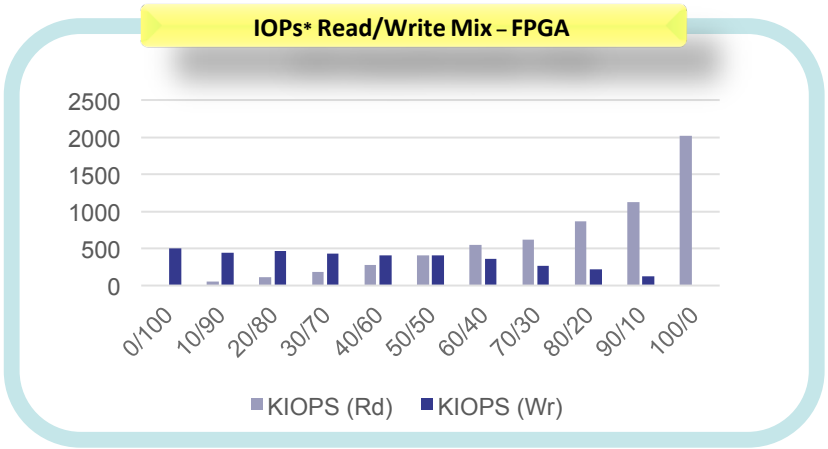
- ETRNIC Driver
- NVMe-oF Driver
- Ethernet Driver
- Network Stack
- Management Interfaces & APIs

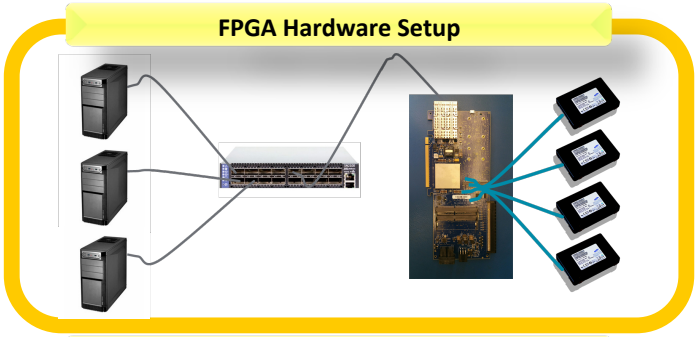
Benchmarking Hardware Topology – FPGA Based Setup



Benchmarking Data: Performance

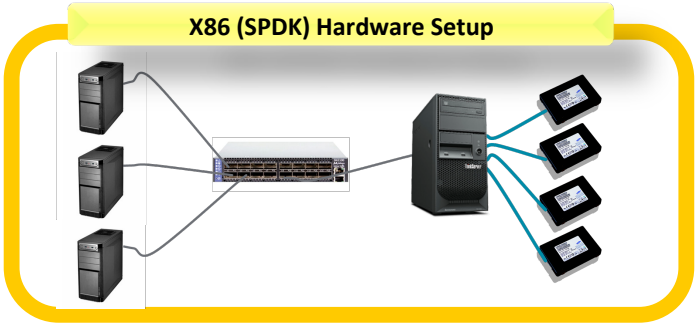






Hardware Configuration

- Initiator: Intel(R) i7-5820K CPU @ 3.30GHz, 12 cores, 8GB RAM
- CNA: Mellanox CX-4
- Switch: MSN2100-CB2F Model MSN2100, 16 QSFP28 ports
- Target (FPGA): Fidus Sidewinder Board ZU19, 2x100G QSFP+
- Target (x86): Dual Socket E5-2620 (12 cores each), 128GB
- SSD: PCIe Gen3 x4, Capable of 740k Read IOPs each



*Data generated using fio and block_size=4k

Benchmarking Data: Latency

SPDK:

```
read : io=2462.8MB, bw=42031KB/s, iops=10507, runt= 60000msec
slat (usec): min=1, max=35, avg= 2.10, stdev= 0.51
clat (usec): min=67, max=2161, avg=92.29, stdev=11.75
lat (usec): min=75, max=2163, avg=94.47, stdev=11.7
```

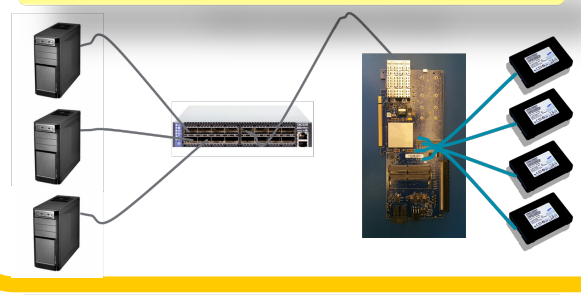
FPGA:

```
read : io=1219.9MB, bw=41637KB/s, iops=10409, runt= 30001msec
slat (usec): min=1, max=70, avg= 3.37, stdev= 2.65
clat (usec): min=45, max=188, avg=91.47, stdev=11.49
lat (usec): min=74, max=194, avg=95.00, stdev=12.07
```

DAS:

```
read : io=1283.1MB, bw=43823KB/s, iops=10955, runt= 30001msec
slat (usec): min=3, max=43, avg= 3.73, stdev= 0.62
clat (usec): min=46, max=2209, avg=86.18, stdev=13.02
lat (usec): min=72, max=2213, avg=90.08, stdev=13.02
```

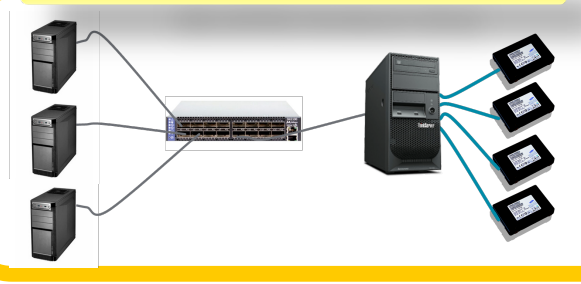
FPGA Hardware Setup



Hardware Configuration

- Initiator: Intel(R) i7-5820K CPU @ 3.30GHz, 12 cores, 8GB RAM
- CNA: Mellanox CX-4
- Switch: MSN2100-CB2F Model MSN2100, 16 QSFP28 ports
- Target (FPGA): Fidus Sidewinder Board ZU19, 2x100G QSFP+
- Target (x86): Dual Socket E5-2620 (12 cores each), 128GB
- SSD: PCIe Gen3 x4, Capable of 740k Read IOPs each

X86 (SPDK) Hardware Setup



*Data generated using fio and block_size=4k

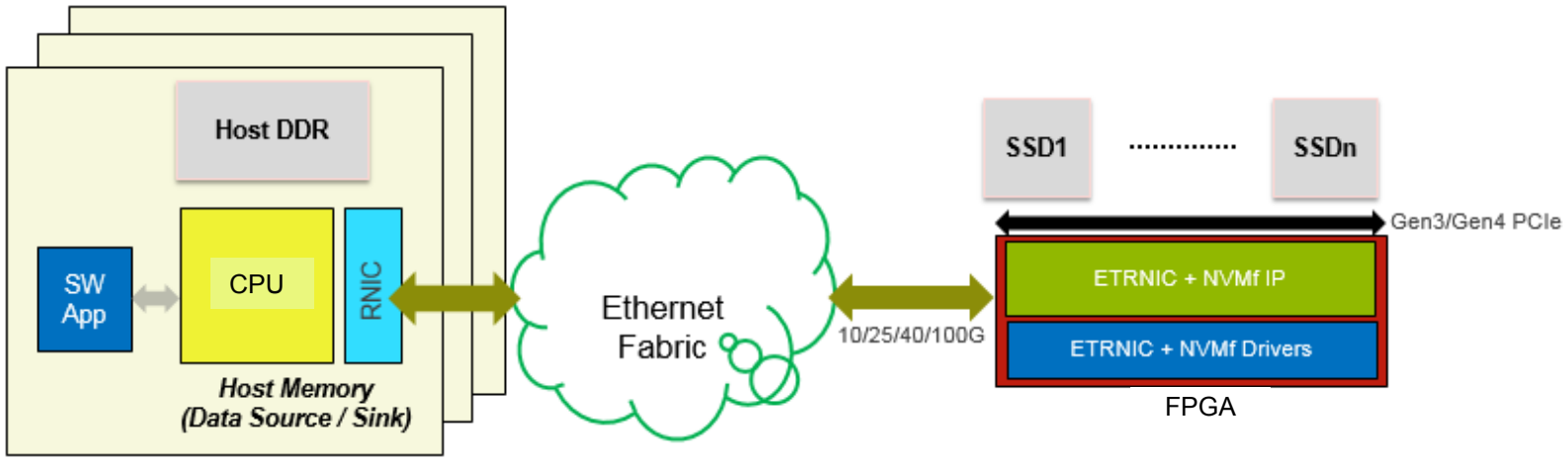
Conclusions

- FPGA Architectures today provide a highly integrated solution to NVMe over Fabric acceleration problem
- Allows for low latency, high throughput storage architectures with built-in accelerators
- FPGAs allow for workload specific optimizations enabled with partial reconfiguration

Thank you!

BACKUP

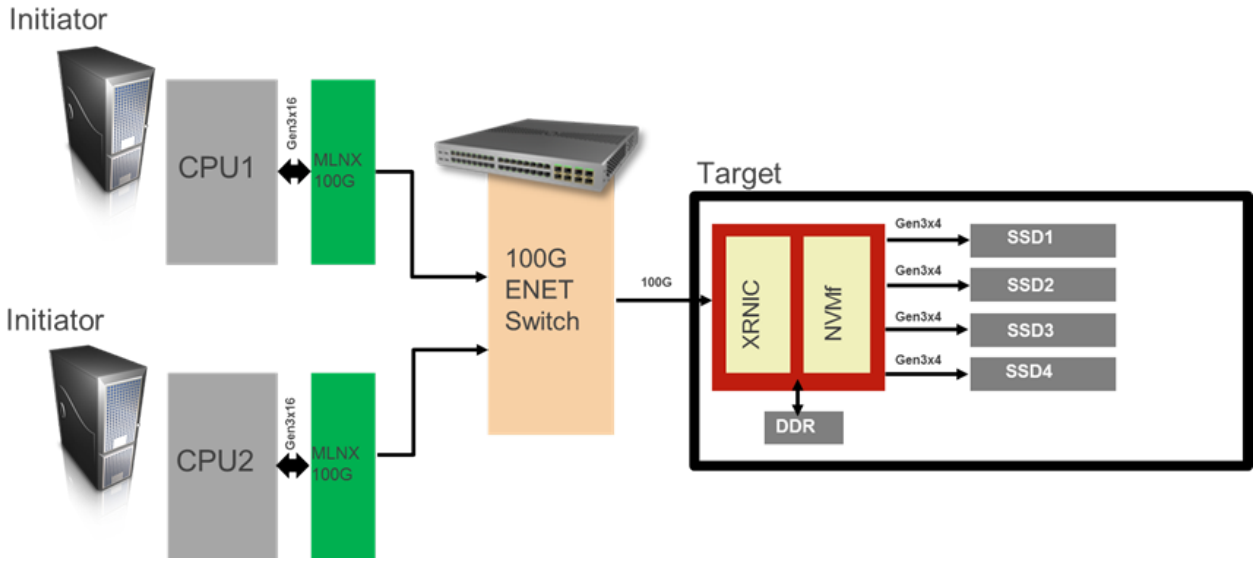
NVMe-oF Integrated Architecture



- Integrated end-point RNIC
  - HW implementation for Target/end point RoCE v2
  - Supports up to 128 Initiators
- Completely HW accelerated
  - Doorbell exchanges through HW handshake
- No P2P bottleneck
- All end point drivers running on FPGA processor

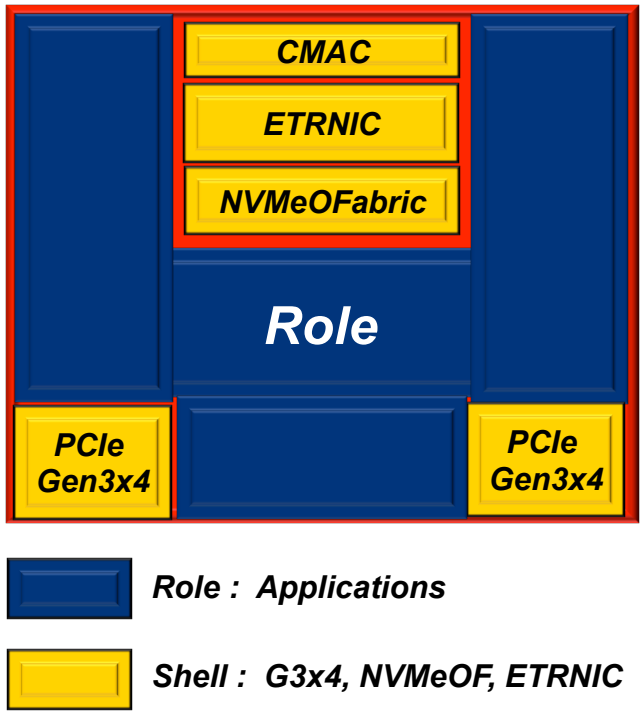
Xilinx ETRNIC + NVMf Reference Design

- Demo configuration
  - Fidus Sideweinder board with **ZU19EG**
  - 4 Samsung SSDs, directly attached to FPGA via PCIe Gen3x4
  - 2 Initiators running Open NVMf
  - 100G Mellanox RNIC on the initiators



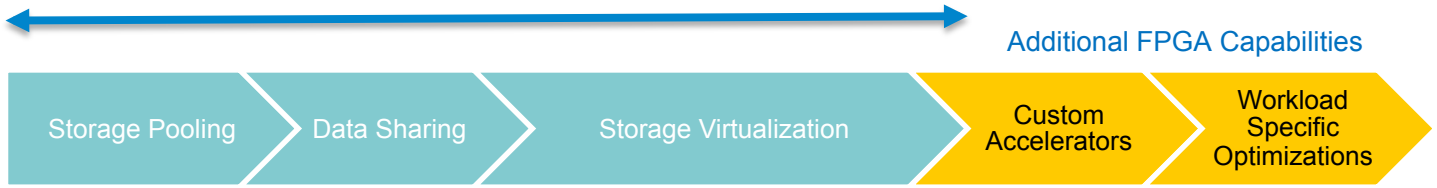
NVMf + ETRNIC : FPGA Value Add

- NVMf + ETRNIC solution
- Shell and Role Model of FPGA
- Shell FPGA : Always “ON”
  - Hard cores : PCIeG3x8
  - Soft cores : NVMeOFabric, ETRNIC
- Role FPGA
  - Applications can be ported : FPGA
  - E2E datapath protection/correction
  - End point Security
  - Caching algorithms etc



FPGA Value Add

Attribute	Xeon CPU + External NIC	Integrated NIC + CPU	FPGA
Solution Cost	High. Xeon adds cost	Lower	Lower
Solution Power	Higher	Lower	Lower
Latency, Latency Outliers	High Latency, large variation in latency	Low latency. Shared data and control plane within the same CPU increases latency variation	Low Latency. Orthogonal data and control place minimize latency variation
Custom Accelerators	In Software	In Software. Limited by integrated processor capability	In Hardware
Hardware Reconfiguration	No	No	Yes



Flash Memory Summit

Xilinx Device options: MPSoC

Zynq® UltraScale+™ MPSoCs: EG Devices

		Device Name[1]	ZU2EG	ZU3EG	ZU4EG	ZU5EG	ZU6EG	ZU7EG	ZU9EG	ZU11EG	ZU15EG	ZU17EG	ZU19EG
Processing System (PS)	Application Processor Unit	Processor Core	**Quad-core** ARM® Cortex™-A53 MPCore™ up to 1.5GHz										
		Memory w/ECC	L1 Cache 32KB I / D per core, L2 Cache 1MB, on-chip Memory 256KB										
	Real-Time Processor Unit	Processor Core	**Dual-core** ARM Cortex-R5 MPCore™ up to 600MHz										
		Memory w/ECC	L1 Cache 32KB I / D per core, Tightly Coupled Memory 128KB per core										
	Graphic & Video Acceleration	Graphics Processing Unit	Mali™-400 MP2 up to 667MHz										
		Memory	L2 Cache 64KB										
	External Memory	Dynamic Memory Interface	x32/x64: DDR4, LPDDR4, DDR3, DDR3L, LPDDR3 with ECC										
		Static Memory Interfaces	NAND, 2x Quad-SPI										
	Connectivity	High-Speed Connectivity	PCIe® Gen2 x4, 2x USB3.0, SATA 3.1, DisplayPort, 4x Tri-mode Gigabit Ethernet										
		General Connectivity	2xUSB 2.0, 2x SD/SDIO, 2x UART, 2x CAN 2.0B, 2x I2C, 2x SPI, 4x 32b GPIO										
	Integrated Block Functionality	Power Management	Full / Low / PL / Battery Power Domains										
		Security	RSA, AES, and SHA										
		AMS - System Monitor	10-bit, 1MSPS - Temperature, Voltage, and Current Monitor										
PS to PL Interface			12 x 32/64/128b AXI Ports										
Programmable Logic (PL)	Programmable Functionality	System Logic Cells (K)	103	154	192	256	469	504	600	653	747	926	1,143
		CLB Flip-Flops (K)	94	141	176	234	429	461	548	597	682	847	1,045
		CLB LUTs (K)	47	71	88	117	215	230	274	299	341	423	523
	Memory	Max. Distributed RAM (Mb)	1.2	1.8	2.6	3.5	6.9	6.2	8.8	9.1	11.3	8.0	9.8
		Total Block RAM (Mb)	5.3	7.6	4.5	5.1	25.1	11.0	32.1	21.1	26.2	28.0	34.6
		UltraRAM (Mb)	-	-	14.0	18.0	-	27.0	-	22.5	31.5	28.7	36.0
	Clocking	Clock Management Tiles (CMTs)	3	3	4	4	4	8	4	8	4	11	11
	Integrated IP	DSP Slices	240	360	728	1,248	1,973	1,728	2,520	2,928	3,528	1,590	1,968
		PCI Express® Gen 3x16 / Gen4x8	-	-	2	2	-	2	-	4	-	4	5
		150G Interlaken	-	-	-	-	-	-	-	1	-	2	4
		100G Ethernet MAC/PCS w/RS-FEC	-	-	-	-	-	-	-	2	-	2	4
		AMS - System Monitor	1	1	1	1	1	1	1	1	1	1	1
	Transceivers	GTH 16.3Gb/s Transceivers	-	-	16	16	24	24	24	32	24	44	44
		GTY 32.75Gb/s Transceivers	-	-	-	-	-	-	-	16	-	28	28
	Speed Grades	Extended[2]	-1 -2 -2L			-1 -2 -2L -3				-1 -2 -2L -3			
		Industrial	-1 -1L -2										

Flash Memory Summit

Xilinx Device options: Virtex US+

Device Name	VU3P	VU5P	VU7P	VU9P	VU11P	VU13P	VU31P	VU33P	VU35P	VU37P
System Logic Cells (K)	862	1,314	1,724	2,586	2,835	3,780	962	962	1,907	2,852
CLB Flip-Flops (K)	788	1,201	1,576	2,364	2,592	3,456	879	879	1,743	2,607
CLB LUTs (K)	394	601	788	1,182	1,296	1,728	440	440	872	1,304
Max. Distributed RAM (Mb)	12.0	18.3	24.1	36.1	36.2	48.3	12.5	12.5	24.6	36.7
Total Block RAM (Mb)	25.3	36.0	50.6	75.9	70.9	94.5	23.6	23.6	47.3	70.9
UltraRAM (Mb)	90.0	132.2	180.0	270.0	270.0	360.0	90.0	90.0	180.0	270.0
HBM DRAM (GB)	–	–	–	–	–	–	4	8	8	8
HBM AXI Interfaces	–	–	–	–	–	–	32	32	32	32
Clock Mgmt Tiles (CMTs)	10	20	20	30	12	16	4	4	8	12
DSP Slices	2,280	3,474	4,560	6,840	9,216	12,288	2,880	2,880	5,952	9,024
Peak INT8 DSP (TOP/s)	7.1	10.8	14.2	21.3	28.7	38.3	8.9	8.9	18.6	28.1
PCIe® Gen3 x16 / Gen4 x8	2	4	4	6	3	4	4	4	5	6
CCIX Ports[1]	–	–	–	–	–	–	4	4	4	4
150G Interlaken	3	4	6	9	6	8	0	0	2	4
100G Ethernet w/ RS-FEC	3	4	6	9	9	12	2	2	5	8
Max. Single-Ended HP I/Os	520	832	832	832	624	832	208	208	416	624
GTY 32.75Gb/s Transceivers	40	80	80	120	96	128	32	32	64	96
Extended[2]	-1 -2 -2L -3	-1 -2 -2L -3	-1 -2 -2L -3	-1 -2 -2L -3	-1 -2 -2L -3	-1 -2 -2L -3	-1 -2 -2L -3	-1 -2 -2L -3	-1 -2 -2L -3	-1 -2 -2L -3
Industrial	-1 -2	-1 -2	-1 -2	-1 -2	-1 -2	-1 -2	–	–	–	–

Footprint compatible with 20nm UltraScale Devices with same footprint identifier

Footprint[3,4]	Dimensions (mm)	HP I/O, GTY 32.75Gb/s								
C1517	40x40	520, 40								
F1924[5]	45x45					624, 64				
A2104	47.5x47.5		832, 52	832, 52	832, 52					
	52.5x52.5[6]						832, 52			
B2104	47.5x47.5		702, 76	702, 76	702, 76	572, 76				
	52.5x52.5[6]						702, 76			
C2104	47.5x47.5		416, 80	416, 80	416, 104	416, 96				
	52.5x52.5[6]						416, 104			
D2104	47.5x47.5				676, 76	572, 76				
	52.5x52.5[6]						676, 76			
A2577	52.5x52.5				448, 120	448, 96	448, 128			
H1924	45x45							208, 32		
H2104	47.5x47.5								208, 32	416, 64
H2892	55x55									416, 64	624, 96

Virtex® UltraScale+™ FPGAs