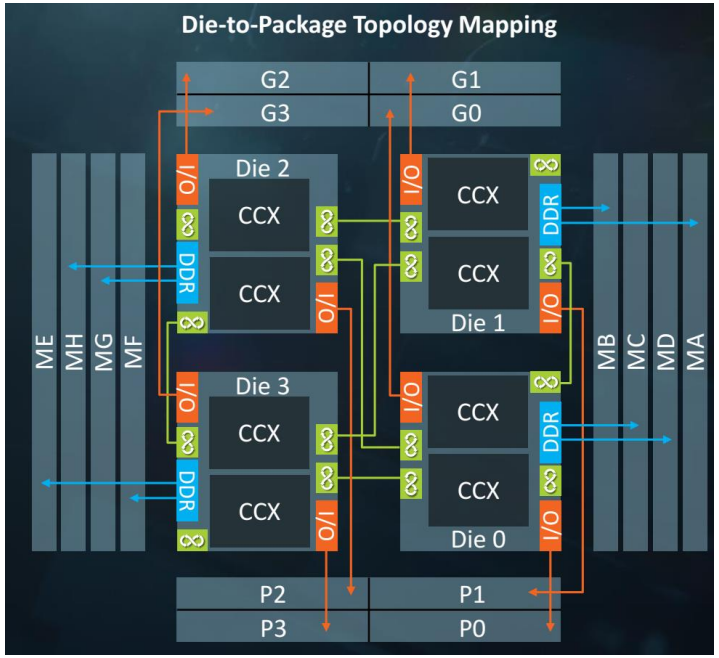
Achieve stable high performance DPDK Application on modern CPU

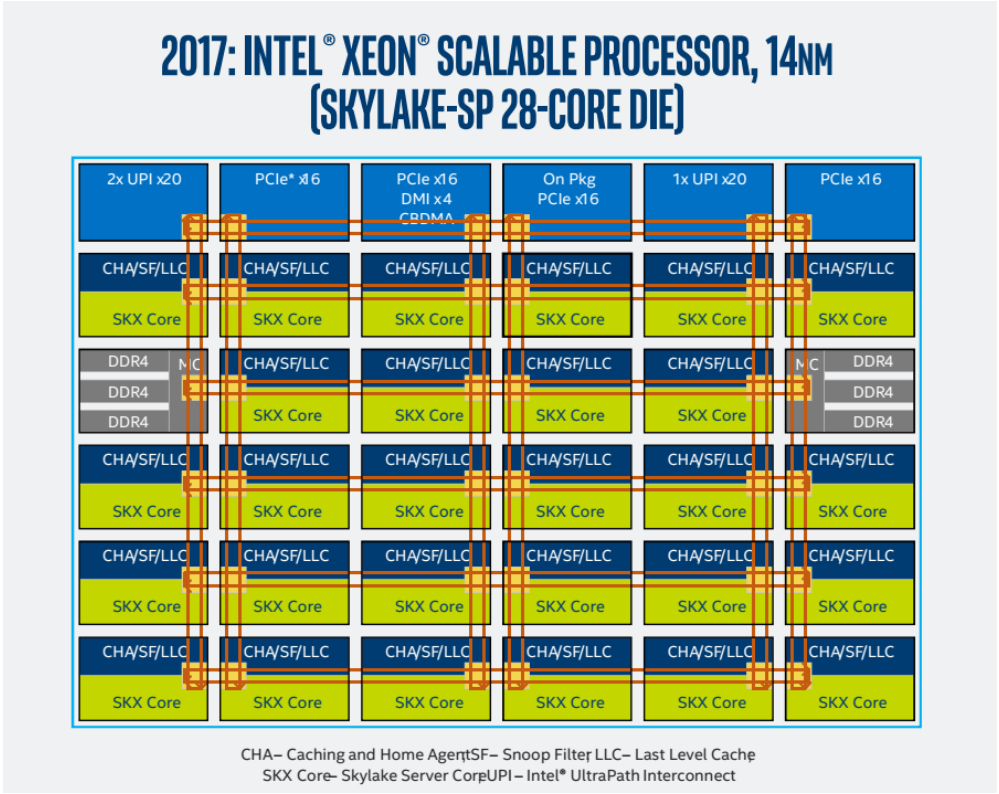
TAO YANG
XUEKUN HU
INTEL

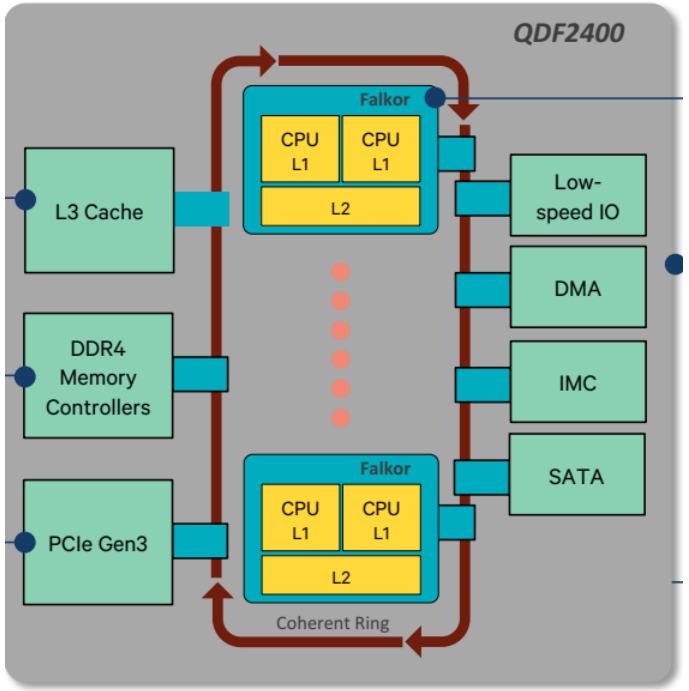
Agenda

1. Modern CPU Architecture
2. Performance impact on shared resource
    1. Shared EU and L1/L2 Cache
    2. Shared L3 Cache
    3. Shared Core Power
3. Summary

Server CPU in Hot Chips 2017







- https://www.hotchips.org/wp-content/uploads/hc_archives/hc29/HC29.22-Tuesday-Pub/HC29.22.90-Server-Pub/HC29.22.921-EPYC-Lepak-AMD-v2.pdf
- https://www.hotchips.org/wp-content/uploads/hc_archives/hc29/HC29.22-Tuesday-Pub/HC29.22.90-Server-Pub/HC29.22.930-Xeon-Skylake-sp-Kumar-Intel.pdf
- https://www.hotchips.org/wp-content/uploads/hc_archives/hc29/HC29.22-Tuesday-Pub/HC29.22.90-Server-Pub/HC29.22.942-Centriq-2400-Wolford-Qualcomm%20Final%20Submission%20corrected.pdf

Shared Execution Engine and L1/L2 Cache

DPDK
DATA PLANE DEVELOPMENT KIT

- Hyper-Threading Technology

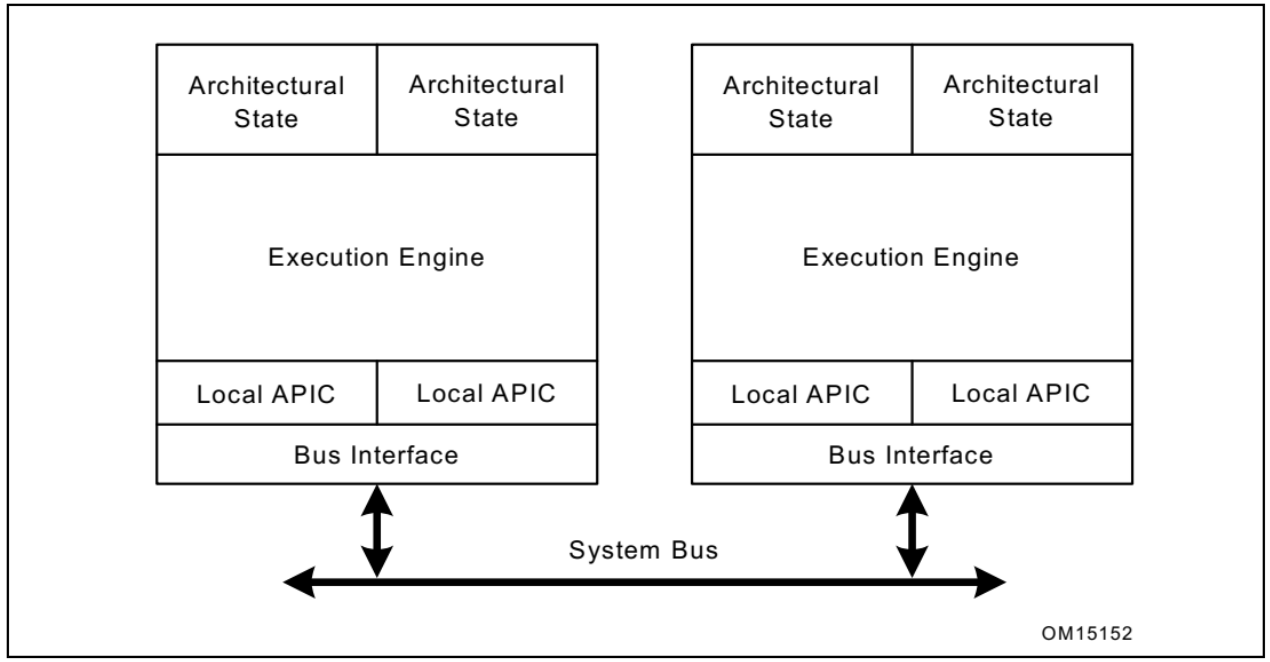


Figure 2-16. Hyper-Threading Technology on an SMP

- **Intel® 64 and IA-32 Architectures Software Developer Manuals**

Linux tool for hyper-thread

DPDK
DATA PLANE DEVELOPMENT KIT

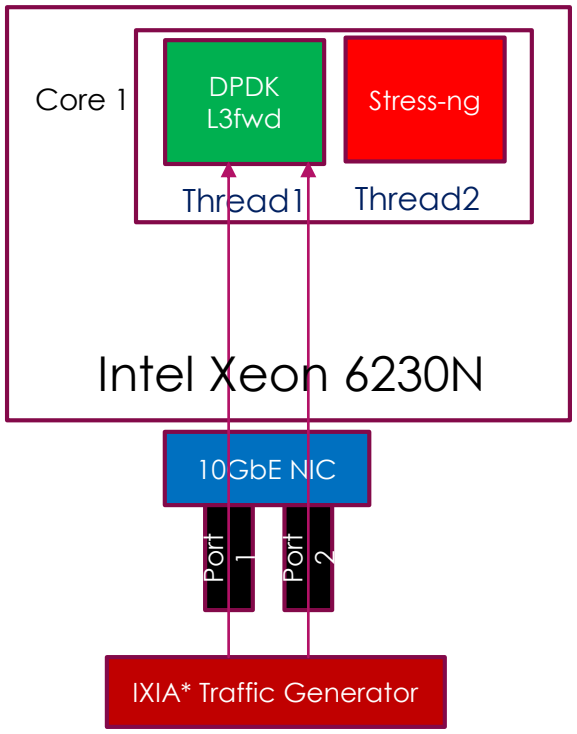
- Hyper-thread and cores in the system

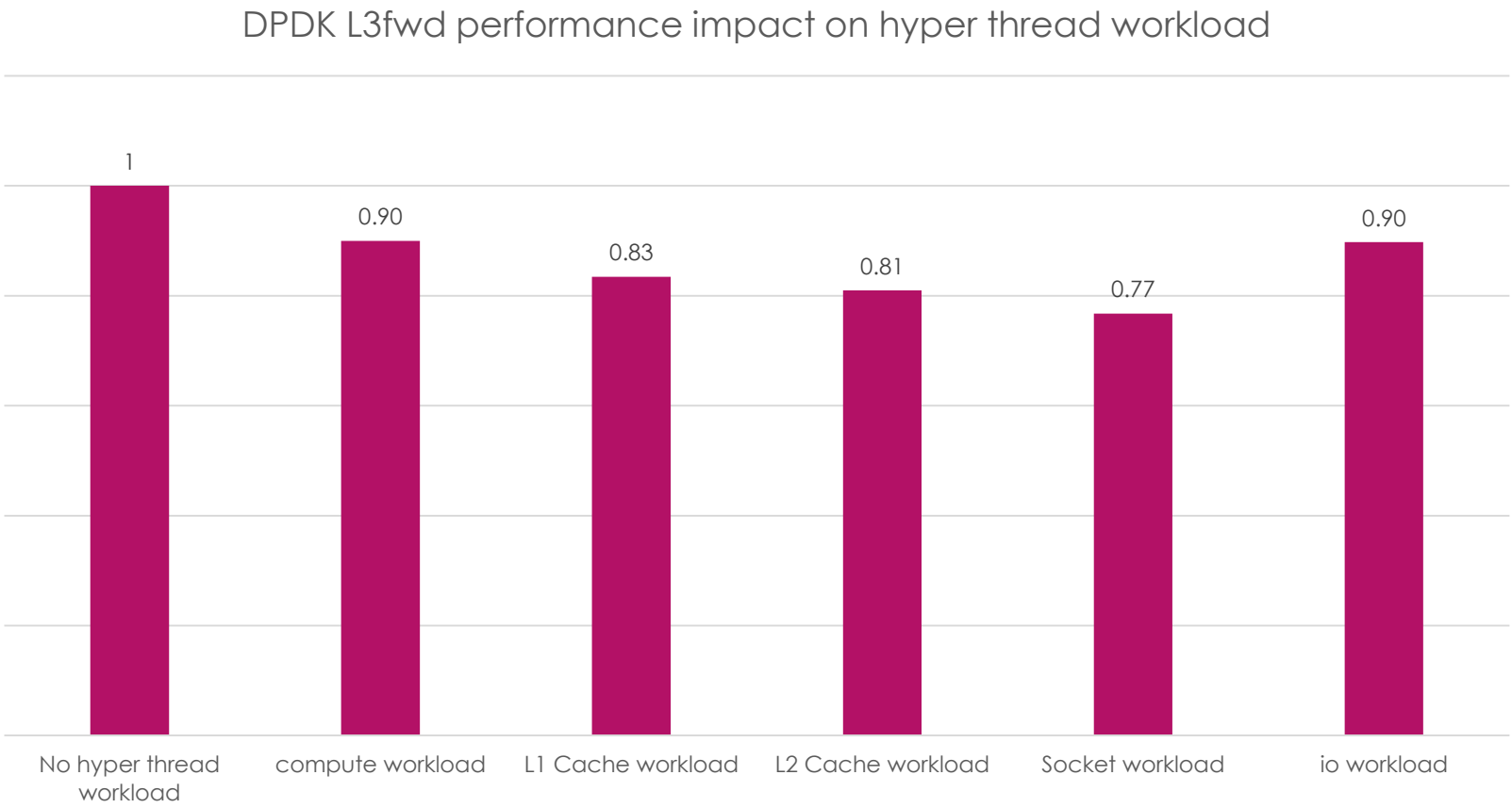
```
[root@wolfpass-6230n ~]# lscpu
CPU(s):                80
Thread(s) per core:    2
Core(s) per socket:    20
Socket(s):             2
NUMA node(s):          2

[root@wolfpass-6230n ~]# lscpu -e
CPU NODE SOCKET CORE L1d:L1i:L2:L3 ONLINE MAXMHZ    MINMHZ
0   0    0      0    0:0:0:0       yes    3900.0000 800.0000
1   0    0      1    1:1:1:0       yes    3900.0000 800.0000
.....
40  0    0      0    0:0:0:0       yes    3900.0000 800.0000
.....
```

- Binding application to core
- taskset -c CORE-ID DPDK-APP

Hyper-thread performance impact

- DPDK L3fwd on 1 core/1 Thread (Intel Xeon 6230N 2.30GHz) with 2*10G port
- Stress workload running on the other hyper thread





Shared L3 Cache

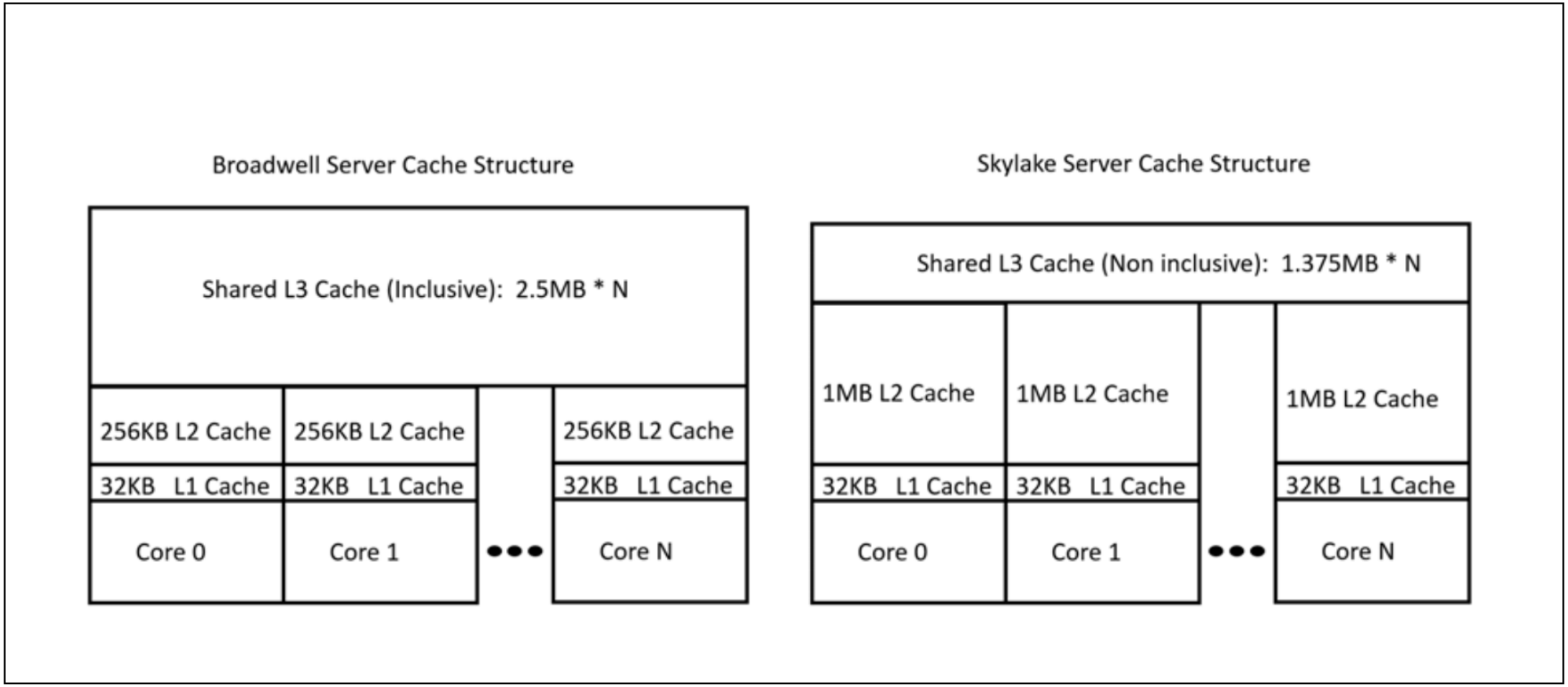
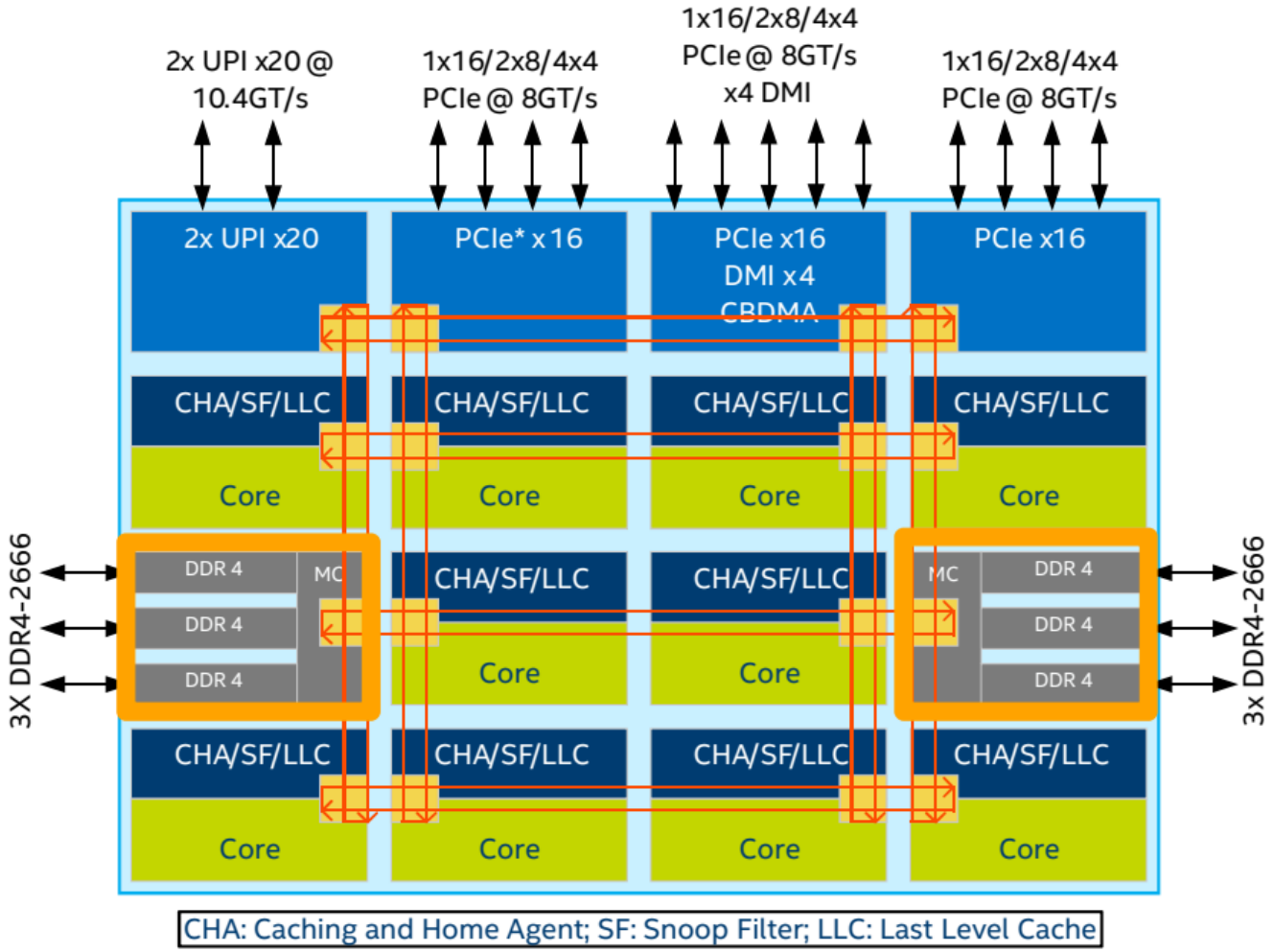


Figure 2-2. Broadwell Microarchitecture and Skylake Server Microarchitecture Cache Structures

- Intel® 64 and IA-32 architectures optimization reference manual

Shared Memory Bandwidth



CHA: Caching and Home Agent; SF: Snoop Filter; LLC: Last Level Cache

- https://www.hotchips.org/wp-content/uploads/hc_archives/hc29/HC29.22-Tuesday-Pub/HC29.22.90-Server-Pub/HC29.22.930-Xeon-Skylake-sp-Kumar-Intel.pdf

Intel® Resource Director Technology

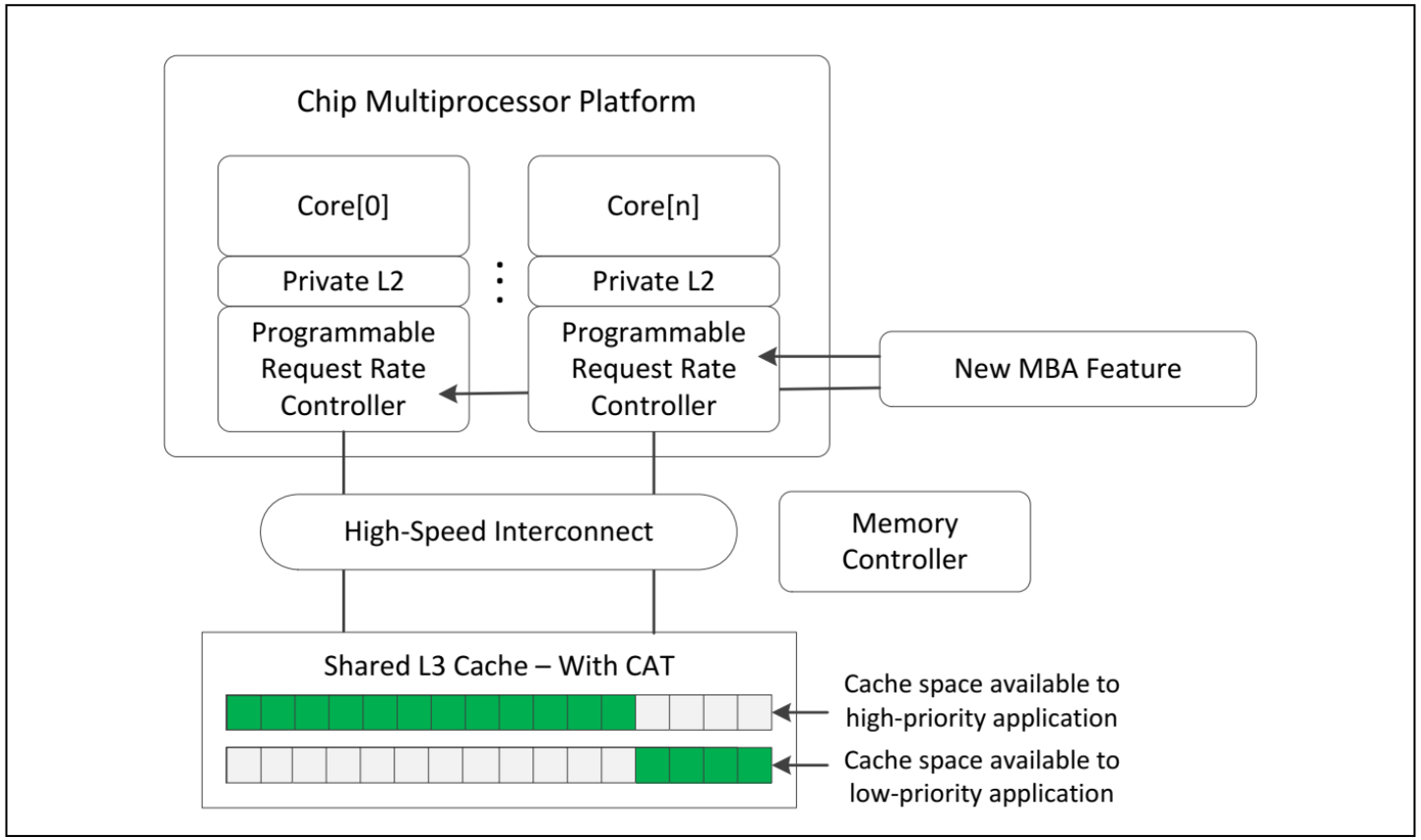


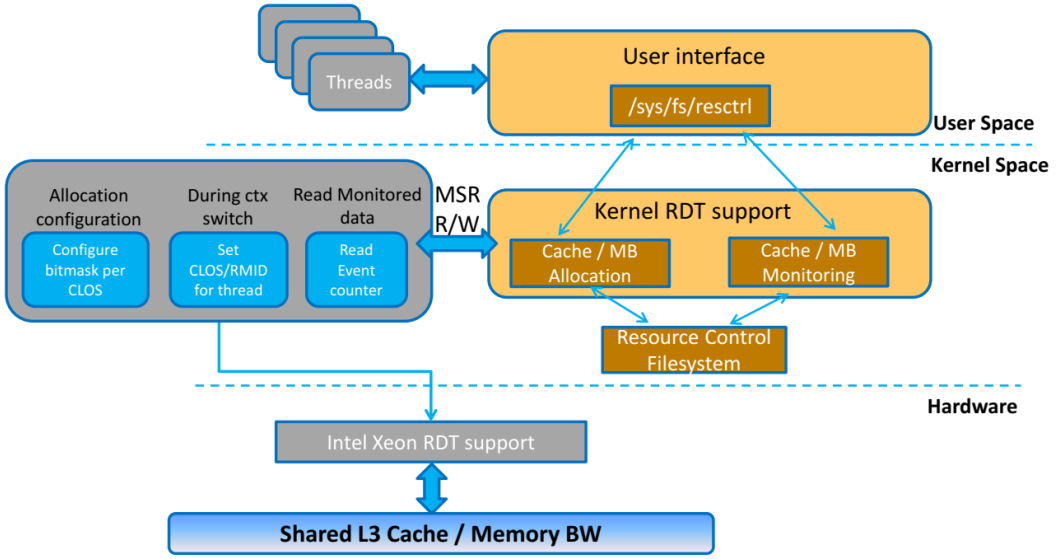
Figure 17-38. A High-Level Overview of the MBA Feature

- Intel® 64 and IA-32 Architectures Software Developer Manuals
- https://www.intel.com/content/www/us/en/architecture-and-technology/resource-director-technology.html

Linux Tools for Resource Control

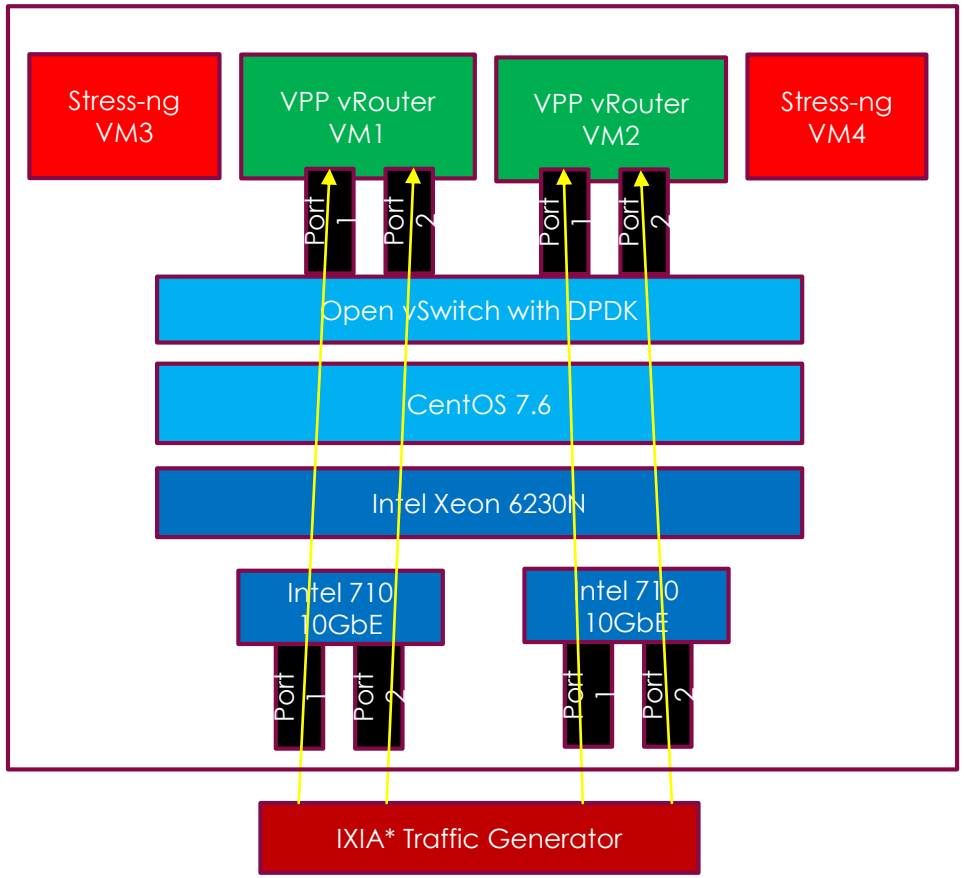
DPDK
DATA PLANE DEVELOPMENT KIT

- Intel RDT Kernel Interface Documentation
  https://www.kernel.org/doc/html/latest/x86/resctrl_ui.html



- Intel RDT Reference Software Package （Direct access CPU register）
  https://github.com/intel/intel-cmt-cat

RDT Test configuration



Physical Core	Process/VM	"CAT with Aggressors" Case		Cache Allocation Scheme											Memory Bandwidth Allocation Scheme
		CoS	Capacity Bit Mask (CBM)	11 bit CBM representation											
33,35,37 60,73,75,77	Other App	3	0xC	10	9	8	7	6	5	4	3	2	1	0	10%
20	ovs-vswitchd	3	0xC	10	9	8	7	6	5	4	3	2	1	0	10%
21,26,27, 28,29,36	OVS-DPDK PMD	1	0x7F0	10	9	8	7	6	5	4	3	2	1	0	100%
22,23,24	VM1 - SUT	1	0x7F0	10	9	8	7	6	5	4	3	2	1	0	100%
30,31,32	VM2 - SUT	1	0x7F0	10	9	8	7	6	5	4	3	2	1	0	100%
25,34,65,74	VM3 - Noisy Neighbor	2	0x3	10	9	8	7	6	5	4	3	2	1	0	10%
38,39,78,79	VM4 - Noisy Neighbor	2	0x3	10	9	8	7	6	5	4	3	2	1	0	10%
0-19,40-59	OS on CPU 0	0	0x7FF												100%

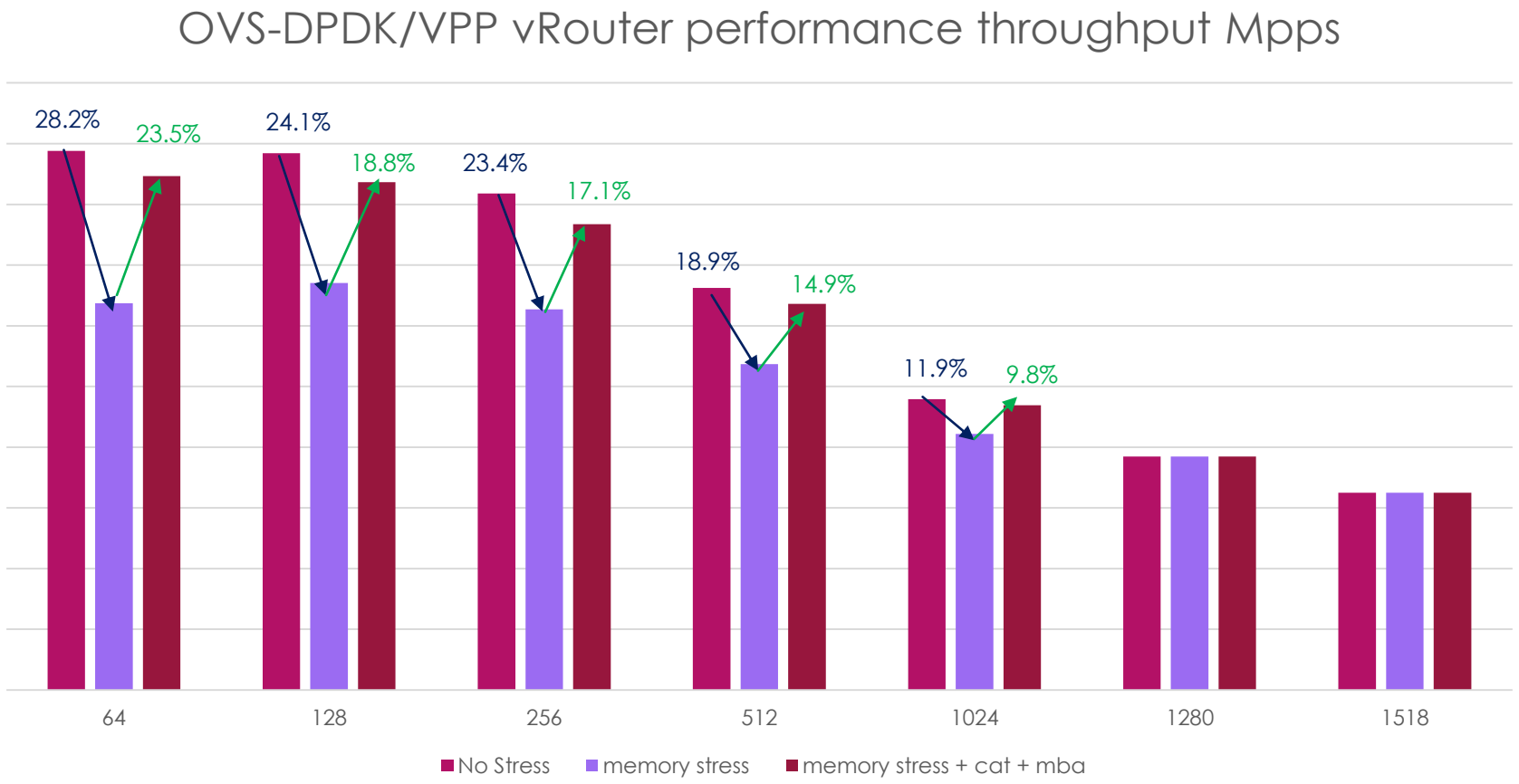
```
pqos -e "llc:0=0x7ff;llc:1=0x7f0;llc:2=0x3;llc:3=0xc;"
pqos -e "mba:0=100;mba:1=100;mba:2=10;mba:3=10;“
pqos -a "llc:0=0-19,40-59"
pqos -a "llc:1=21,26,27,28,29,36,22,23,24,30,31,32;llc:2=25,34,38,39;llc:3=20,33,35,37"
pqos -a "llc:1=61,66,67,68,69,76,62,63,64,70,71,72;llc:2=65,74,78,79;llc:3=60,73,75,77"
```

RDT Test data

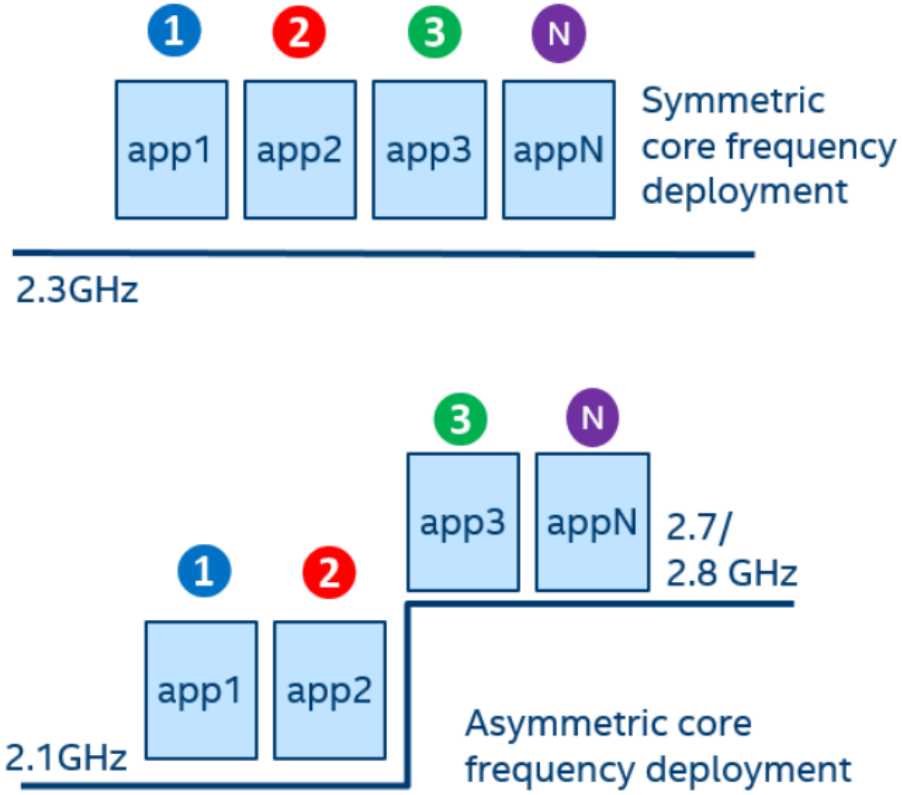
Socket 0		Socket 1	
Memory Channel Monitoring		Memory Channel Monitoring	
Mem Ch 0: Reads (MB/s):	4.29	Mem Ch 0: Reads (MB/s):	6317.62
Writes(MB/s):	0.84	Writes(MB/s):	2499.69
Mem Ch 1: Reads (MB/s):	4.88	Mem Ch 1: Reads (MB/s):	6312.72
Writes(MB/s):	1.24	Writes(MB/s):	2499.57
Mem Ch 2: Reads (MB/s):	4.95	Mem Ch 2: Reads (MB/s):	6311.08
Writes(MB/s):	0.83	Writes(MB/s):	2499.12
Mem Ch 3: Reads (MB/s):	5.72	Mem Ch 3: Reads (MB/s):	6324.11
Writes(MB/s):	1.41	Writes(MB/s):	2499.27
Mem Ch 4: Reads (MB/s):	5.00	Mem Ch 4: Reads (MB/s):	6319.79
Writes(MB/s):	0.85	Writes(MB/s):	2499.24
Mem Ch 5: Reads (MB/s):	4.37	Mem Ch 5: Reads (MB/s):	6319.52
Writes(MB/s):	0.94	Writes(MB/s):	2500.65
NODE 0 Mem Read (MB/s) :	29.21	NODE 1 Mem Read (MB/s) :	37904.83
NODE 0 Mem Write(MB/s) :	6.12	NODE 1 Mem Write(MB/s) :	14997.54
NODE 0 P. Write (T/s):	18711	NODE 1 P. Write (T/s):	9374863
NODE 0 Memory (MB/s):	35.33	NODE 1 Memory (MB/s):	52902.37
System Read Throughput(MB/s):	37934.04		
System Write Throughput(MB/s):	15003.66		
System Memory Throughput(MB/s):	52937.70		

CORE	IPC	MISSES	LLC[KB]	MBL[MB/s]	MBR[MB/s]
20	0.25	10k	0.0	0.3	0.0
21	1.45	3077k	880.0	254.1	0.1
22	0.26	119k	0.0	1.9	0.3
23	1.37	5155k	800.0	333.5	0.0
24	1.37	5179k	640.0	362.5	0.0
25	0.65	142568k	4400.0	13242.8	0.1
26	1.44	3597k	480.0	286.2	0.0
27	1.42	3208k	320.0	311.3	0.1
28	1.28	3305k	480.0	259.5	0.0
29	1.15	16k	0.0	0.4	0.9
30	0.26	165k	80.0	22.9	0.1
31	1.36	5173k	1360.0	356.6	0.0
32	1.46	396k	80.0	10.8	0.0
33	0.25	3k	0.0	0.0	0.0
34	0.65	143417k	4560.0	13504.5	0.2
35	0.25	3k	0.0	0.0	0.0
36	1.15	14k	80.0	0.4	1.3
37	0.25	5k	0.0	0.0	0.0
38	0.66	144511k	5120.0	13468.5	0.0
39	0.65	142900k	1440.0	12969.0	0.0
60	0.25	13k	80.0	2.3	0.0
61	1.11	3816k	640.0	399.9	0.2
62	0.25	5k	0.0	0.0	0.1
63	0.23	9k	0.0	0.9	0.0
64	0.23	6k	0.0	0.8	0.0
65	0.24	22k	0.0	5.5	0.0
66	1.15	3620k	720.0	373.1	0.0
67	0.96	3904k	960.0	283.4	0.0
68	1.13	3654k	800.0	334.0	0.0
69	1.11	20k	0.0	0.3	0.8
70	0.25	4k	0.0	0.2	0.0
71	0.24	6k	0.0	1.0	0.0
72	0.23	6k	0.0	0.1	0.0
73	0.25	3k	0.0	0.0	0.0
74	0.24	28k	0.0	8.4	0.0
75	0.25	3k	0.0	0.0	0.0
76	1.11	15k	0.0	0.2	0.7
77	0.25	3k	0.0	0.1	0.0
78	0.24	27k	0.0	9.0	0.0
79	0.24	26k	0.0	9.4	0.0

Performance data with RDT



Shared Power



- Intel Speed Select Technology - Base Frequency
- https://builders.intel.com/docs/networkbuilders/intel-speed-select-technology-base-frequency-enhancing-performance.pdf

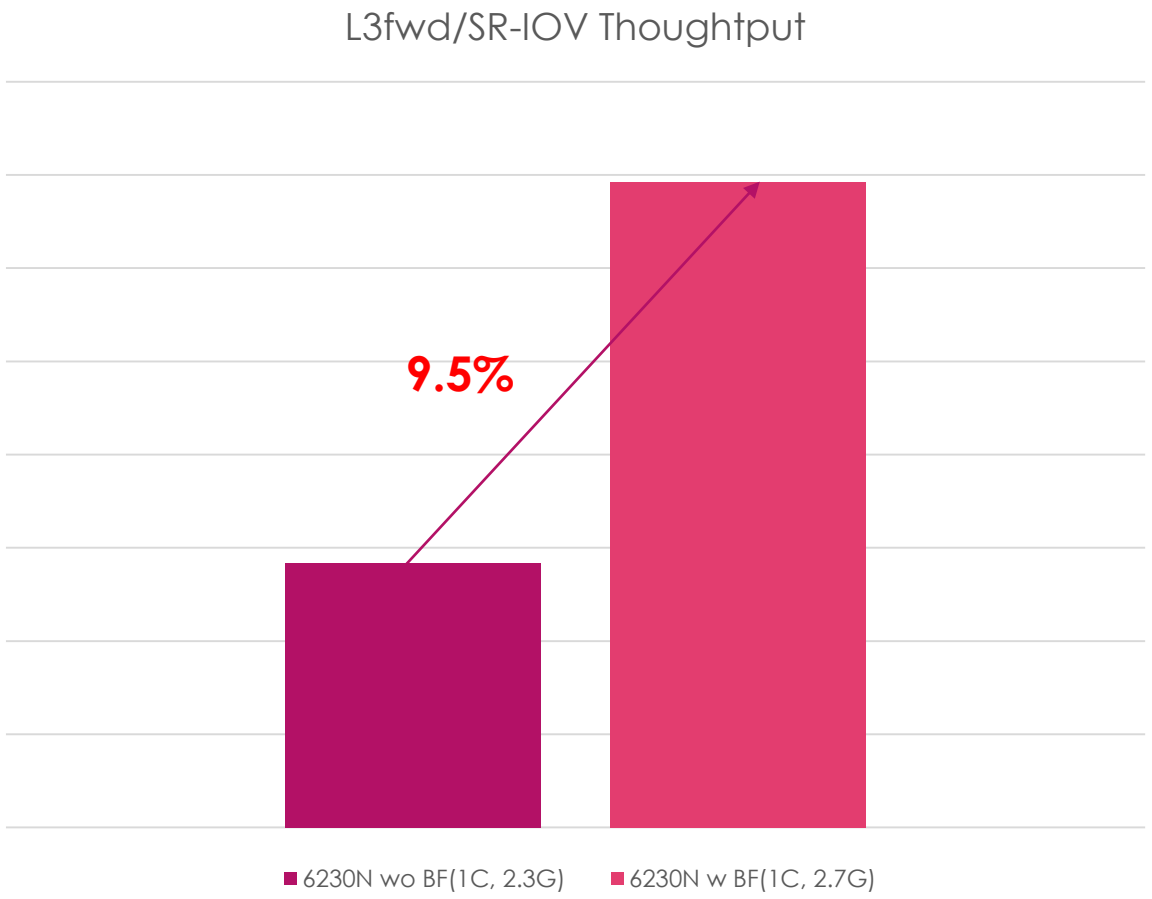
Intel® SST-BF Enabled CPU SKUs

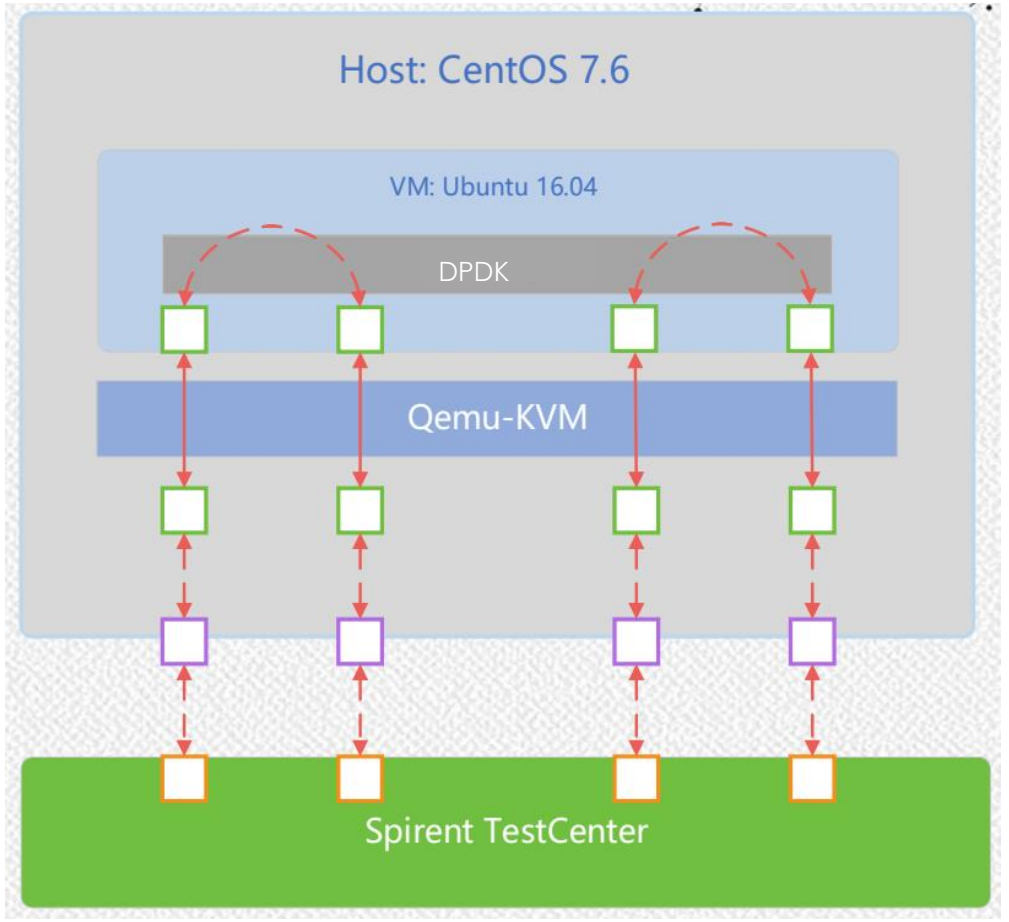
	Base Config		Intel SST-BF Configuration			
	Base Configuration Parameters		Intel SST-BF High Priority Cores		Intel SST-BF Low Priority Cores	
	Cores	SSE Base Freq (GHz)	Cores	SSE Base Freq (GHz)	Cores	SSE Base Freq (GHz)
Intel® Xeon® Gold 6252N Processor	24	2.3	8	2.8	16	2.1
Intel® Xeon® Gold 6230N Processor	20	2.3	6	2.7	14	2.1
Intel® Xeon® Gold 5218N Processor	16	2.3	4	2.7	12	2.1

Linux tool for ISS-BF

- Enable the Intel® SST-BF feature in the BIOS.
- OS can determine high priority cores by enumerating ACPI _CPC object's "guaranteed perf" value for each core for scheduling purposes
  - Linux kernel v5.0.8+ exposes /sys/devices/system/cpu*/cpufreq/base_frequency
- User space script to enable High/Low priority cores
  - https://github.com/intel/CommsPowerManagement

SR-IOV Performance with ISS-BF

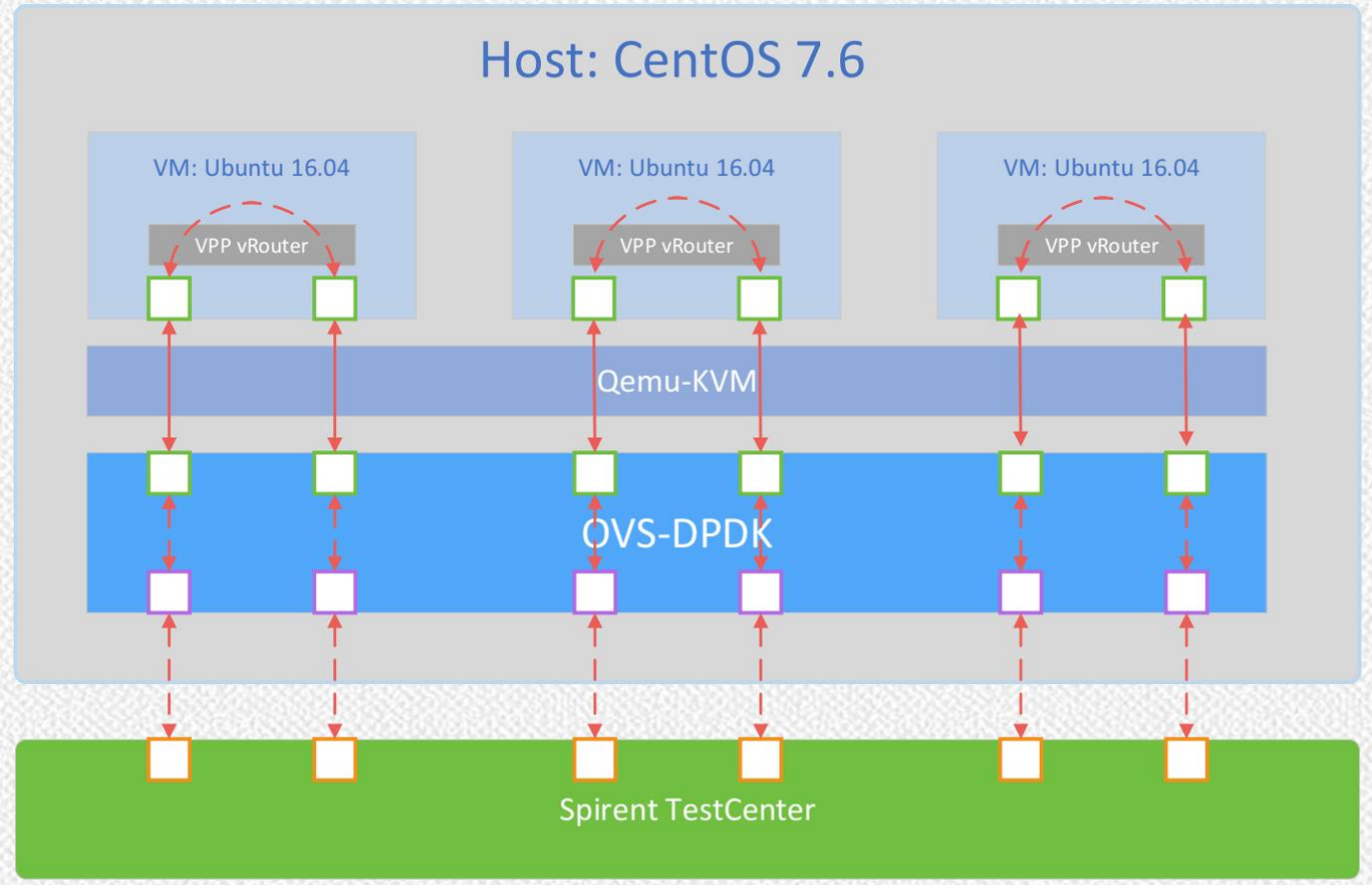




- 1 core in Xeon 6230N, 4 * 10G, SR-IOV passthrough, DPDK l3fwd in VM, packet size 64B

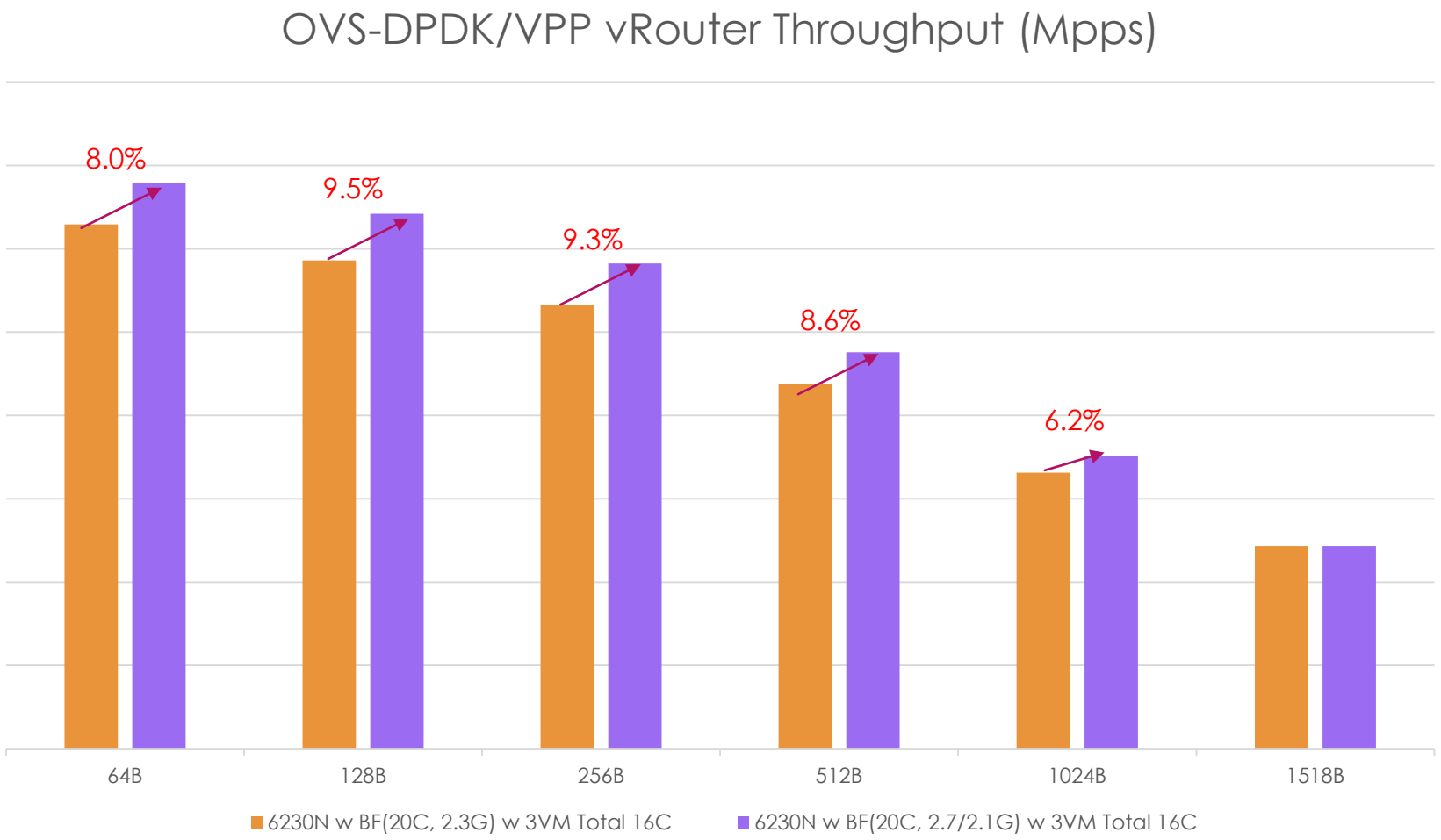
OVS-DPDK Performance with ISS-BF

- 2*Intel Xeon 6230N + 6*10G in a system
- Only used 16 cores in CPU 1
  - 6 cores for OVS-DPDK data plane
    - Low priority Core ID:
      - ovs-vswitchd Core ID: 20
    - High priority Core ID:
      - ovs-pmd: 21,26,27,33,34,36
  - 3 cores for every VPP vRouter VM
    - Low priority Core ID:
      - VM 1: 22,23,24
      - VM 2: 28,29,30
      - VM 3: 32,35,37
- VPP VM core configuration:
  - VM Core 0: for control plane
  - VM Core 1,2: VPP data plane



OVS-DPDK Performance with ISS-BF

- Enable ISS-BF(VPP vRouter 2.1Ghz/OVS-DPDK 2.7Ghz) vs Disable ISS-BF(All Core 2.3Ghz)



Summary

- Many resources are shared in multi-core CPU.
- Application running on different core compete for the shared resources.
- Shared resource partition can reduce the competition and achieve stable high performance.


DPDK
DATA PLANE DEVELOPMENT KIT
Thank You !