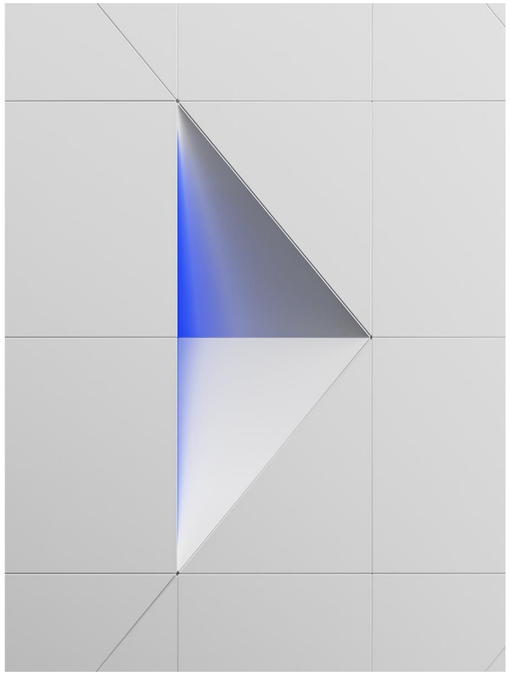
CPU MF Overview and Update for z17

Stephanie DeLuca
IBM Z Performance Specialist
Stephanie.deluca@ibm.com

Contents



Introduction and Overview of CPU MF

Latest CPU MF Formulas and Updates

Value of CPU MF data
- Accurately size IBM Z processors
- Validate achieved performance
- Gain efficiency insights and identify improvements
- Help with performance analysis

Implementation

Introduction to the CPU Measurement Facility (CPU MF)

The CPU Measurement Facility (CPU MF) produces hardware instrumentation data at the logical processor level

– Introduced with IBM z10 and is available today on z10 and later systems

Two modes for collection:

– Counters

- *Track counts of specific CPU activities*
  - How many CPU cycles were used
  - How many instructions completed
  - How often data and instructions were sourced from various levels of cache/memory
  - Counters can vary between machine family

– Sampling

- *Samples that provide snapshot of what the CPU is doing at specific time*

Industry Best Practice and strongly recommended to continuously collect CPU MF counters

IBM Washington Systems Center

Introduction to the CPU Measurement Facility (CPU MF)

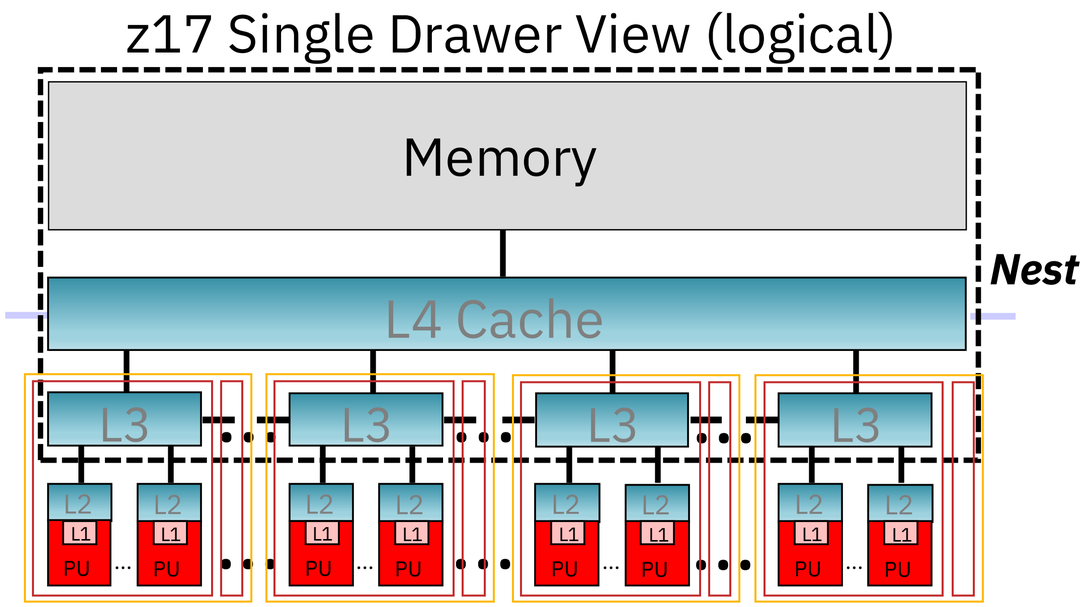
How it works

CPU MF writes counter data to SMF 113 records (z/OS) or Monitor records (z/VM)

– Data can then be used by various performance tooling

CPU MF Formulas are developed from counter data for key metrics in capacity planning and performance analysis

– CPI, workload characterization, use of cache/memory hierarchy



IBM Washington Systems Center

Latest CPU MF Formulas and Updates

CPU MF and LSPR Workload Categories

CPU MF data is used to match customer workloads to one of the LSPR curves to accurately size IBM Z processors

LSPR = Large System Performance Reference

– IBM builds LSPR data from a set of benchmarks running representative workloads

– The unit of capacity is relative "MIPS" and is consistent across all generations of processors

LSPR provides 3 workload curves **to match customer workloads**

– Low
- *Workload curve representing light use of the memory hierarchy*

– Average
- *Workload curve expected to represent the majority of customer workloads*

– High
- *Workload curve representing heavy use of the memory hierarchy*

IBM Capacity Planning Tools use CPU MF data to automatically determine the workload category that should be used for processor sizing

IBM Washington Systems Center

LSPR Workload Category Determination

Updated December 2024

The LSPR Workload Category is determined by two CPU MF metrics:

– L1 Miss Percent (L1MP)

– Relative Nest Intensity (RNI)

LSPR Workload Category Decision Table

L1MP	RNI	LSPR Workload Match
< 3%	>= 0.75 < 0.75	AVERAGE LOW
3% to 6%	>1.0 0.6 to 1.0 < 0.6	HIGH AVERAGE LOW
> 6%	>= 0.75 < 0.75	HIGH AVERAGE

Current table applies to all IBM Z Processors from z10 through z17 CPU MF Data

*** Updated December 2024**

The LSPR Workload Match no longer considers zIIPs

RNI is **not** a Performance metric

– L1MP and RNI allows one to match their workload to an LSPR workload

– Any other use of the RNI metric is not valid

IBM Washington Systems Center

LSPR Workload Category Determination

Updated December 2024

The LSPR Workload Match no longer considers zIIPs

Details

- Workload selection from CPU MF metrics will no longer take zIIP data into account
  - *Previously L1MP and RNI values were calculated for the partition overall, including all data from GCPs and zIIPs*

- Going forward, only GCP data will be used for workload selection
- zIIP values will not affect workload selection for z/VM partitions either

Why?

- From customer data, zIIP data tends to be less nest intensive
- Pushing some customers to LOW or AVG that should have been AVG or HIGH

IBM zPCR updated to reflect this change (December 2024)

IBM Washington Systems Center

CPU MF Formulas for IBM z17

Formulas – z17

Updated April 8, 2025

Note these Formulas may change in the future

Metric	**Calculation –** *note all fields are deltas. SMF113-1s are deltas. SMF 113-2s are cumulative.*
CPI	B0 / B1
PRBSTATE	(P33 / B1) * 100
L1MP	((B2+B4) / B1) * 100
L2P	((E145+E146+E169+E170) / (B2+B4)) * 100
L3P	((E147+E149+E150+E151+E171+E173+E174+E175) / (B2+B4)) * 100
L4LP	((E148+E152+E153+E154+E160+E161+E162+E163+E164+E165+E172+E176+E177+E178) / (B2+B4)) * 100
L4RP	((E155+E166+E167+E168+E179) / (B2+B4)) * 100
MEMP	((E156+E157+E158+E159) / (B2+B4)) * 100
LPARCPU	(((1/CPSP/1,000,000) * B0) / Interval *in Seconds*) * 100

CPI – Cycles per Instruction
Prb State - % Problem State
L1MP – Level 1 Miss Per 100 instructions
L2P – % sourced from Level 2 cache
L3P – % sourced from Level 3 on same Chip cache
L4LP – % sourced from Level 4 Local cache (on same drawer)
L4RP – % sourced from Level 4 Remote cache (on different drawer)
MEMP - % sourced from Memory
LPARCPU - APPL% (GCPs, zIIPs) captured and uncaptured

Workload Characterization
L1 Sourcing from cache/memory hierarchy

B* - Basic Counter Set - Counter Number
P* - Problem-State Counter Set - Counter Number
– See "The Load-Program-Parameter and CPU-Measurement Facilities" SA23-2260 for full description
E* - Extended Counters - Counter Number
– See "IBM The CPU-Measurement Facility Extended Counters Definition for z10, z196/ z114, zEC12 /zBC12, z13/z13s, z14, z15 and z16" SA23-2261-07 for full description
CPSP - SMF113_2_CPSP "CPU Speed"

IBM Washington Systems Center

CPU MF Formulas for IBM z17

Formulas – z17 Additional

Updated April 8, 2025

Note these Formulas may change in the future

Metric	**Calculation–** *note all fields are deltas. SMF113-1s are deltas. SMF 113-2s are cumulative.*
Est Instr Cmplx CPI	CPI – Finite CPI
Finite CPI	E143 / B1
SCPL1M	E143 / (B2+B4)
Rel Nest Intensity	4.7x(0.45xL3P+1.2xL4LP+4.5xL4RP+6.0xMEMP)/100
Eff GHz	CPSP / 1000

Est Instr Cmplx CPI – Estimated Instruction Complexity CPI (infinite L1)
Est Finite CPI – Estimated CPI from Finite cache/memory
Est SCPL1M – Estimated Sourcing Cycles per Level 1 Miss
Rel Nest Intensity –Reflects distribution and latency of sourcing from shared caches and memory
Eff GHz – Effective gigahertz for GCPs, cycles per nanosecond

Workload Characterization
L1 Sourcing from cache/memory hierarchy

B* - Basic Counter Set - Counter Number
P* - Problem-State Counter Set - Counter Number
– See “The Load-Program-Parameter and CPU-Measurement Facilities” SA23-2260 for full description
E* - Extended Counters - Counter Number
– See “IBM The CPU-Measurement Facility Extended Counters Definition for z10, z196/ z114, zEC12 /zBC12, z13/z13s, z14, z15 and z16” SA23-2261-07 for full description
CPSP - SMF113_2_CPSP “CPU Speed”

IBM Washington Systems Center

CPU MF Formulas for IBM z17 Additional AIU

Formulas – z17 Additional AIU

Updated April 8, 2025

Note these Formulas may change in the future

Metric	**Calculation –** *note all fields are deltas. SMF113-1s are deltas. SMF 113-2s are cumulative.*
W_AIU_CPU	(((1/CPSP/1,000,000) * E269) / Interval *in Seconds*) * 100
C_AIU_CPU	(((1/CPSP/1,000,000) * E270) / Interval *in Seconds*) * 100
AIU_CPU	W_AIU_CPU + C_AIU_CPU
LOCAL_AIU %	E272 / E267 * 100
REMOTE_AIU %	E273 / E267 * 100
C_AIU_TIME	E270 / E268 * (1 / CPSP)
W_AIU_TIME	E269 / E268 * (1 / CPSP)

Counter 267 – Increments by one for every NUERAL NETWORK PROCESSING ASSIST instruction executed.

Counter 268 – Increments by one for every NUERAL NETWORK PROCESSING ASSIST instruction executed that ended in Condition Codes 0, 1 or 2.

Counter 269 – Cycles CPU spent obtaining access to IBM Z Integrated Accelerator for AI.

Counter 270 – Cycles CPU is using IBM Z Integrated Accelerator for AI

W_AIU_CPU – Waiting for access to AIU (LPARCPU units)
C_AIU_CPU – Executing on AIU (LPARCPU units)
AIU_CPU – Total AIU CPU (LPARCPU units)
LOCAL_AIU % – Percent of NNPA invocations executing on local AIU
REMOTE_AIU % – Percent of NNPA invocations executing on remote AIU
C_AIU_TIME – AVG microseconds executing per completed instruction
W_AIU_TIME – AVG microseconds waiting per completed instruction

B* - Basic Counter Set - Counter Number
P* - Problem-State Counter Set - Counter Number
– See “The Load-Program-Parameter and CPU-Measurement Facilities” SA23 2260 for full description
E* - Extended Counters - Counter Number
– See “IBM The CPU-Measurement Facility Extended Counters Definition for z10, z196/ z114, zEC12 /zBC12, z13/z13s, z14, z15 and z16” SA23-2261-07 for full description
CPSP - SMF113_2_CPSP “CPU Speed”

IBM Washington Systems Center

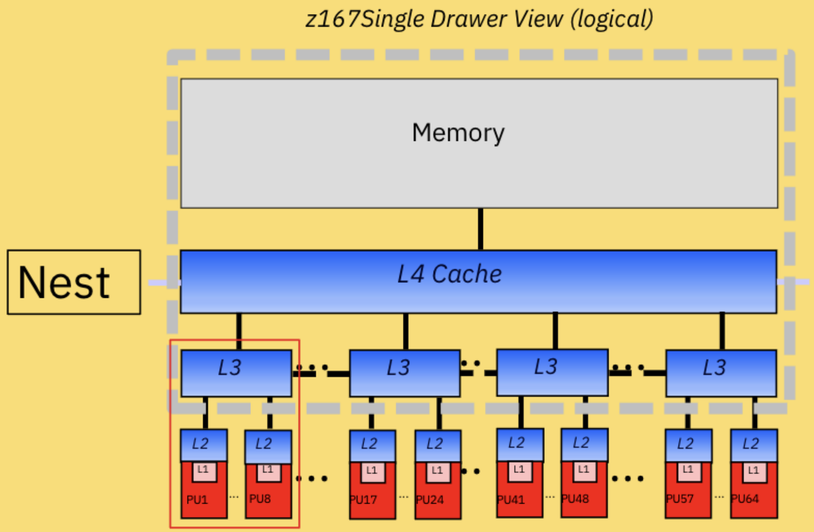
IBM Community Notes

Washington Systems Center

CPU MF IBM Z Best Practice

– z/OS SMF 113 records introduced January 2009 with the IBM System z10

– The IBM Best Practice is to enable the CPU MF records on all IBM Z LPARs



z/OS z17™ Migration "Volunteers"

– Used to validate LSPR Workloads

– Opportunity to ensure your data is used to influence future server design

– **Looking for "Volunteers"**

- Need Before and After Data
- 3 days, 24 hours/day
- SMF 70, 71, 72, 99.14, 113.1 per LPAR
- Production partitions preferred
- Data as close to migration as possible

– No deliverable will be returned

If interested send note to:

stephanie.deluca@ibm.com

Additionally, if you are significantly utilizing your z16 or z17 AIU, we are looking for "Volunteers" – including SMF 30s

Value of CPU MF data

Value of CPU Measurement Facility (CPU MF)

IBM Washington Systems Center

Accurately size IBM Z processors

- IBM recommended method and industry best practice
- Needed on “Before” processor to determine LSPR workload match
- TDA process requires CPU MF Counters to be enabled

Validate achieved IBM Z processor performance

- Needed on “Before” and “After” processors

Gain efficiency insights and identify improvements

- Determine impact of LPAR configuration
- Measure AIU performance
- Identify program inefficiencies (SIIS)
- Understand SMT impact

Help with performance analysis

- Cross-drawer impact
- CPI analysis

Value of CPU Measurement Facility (CPU MF)

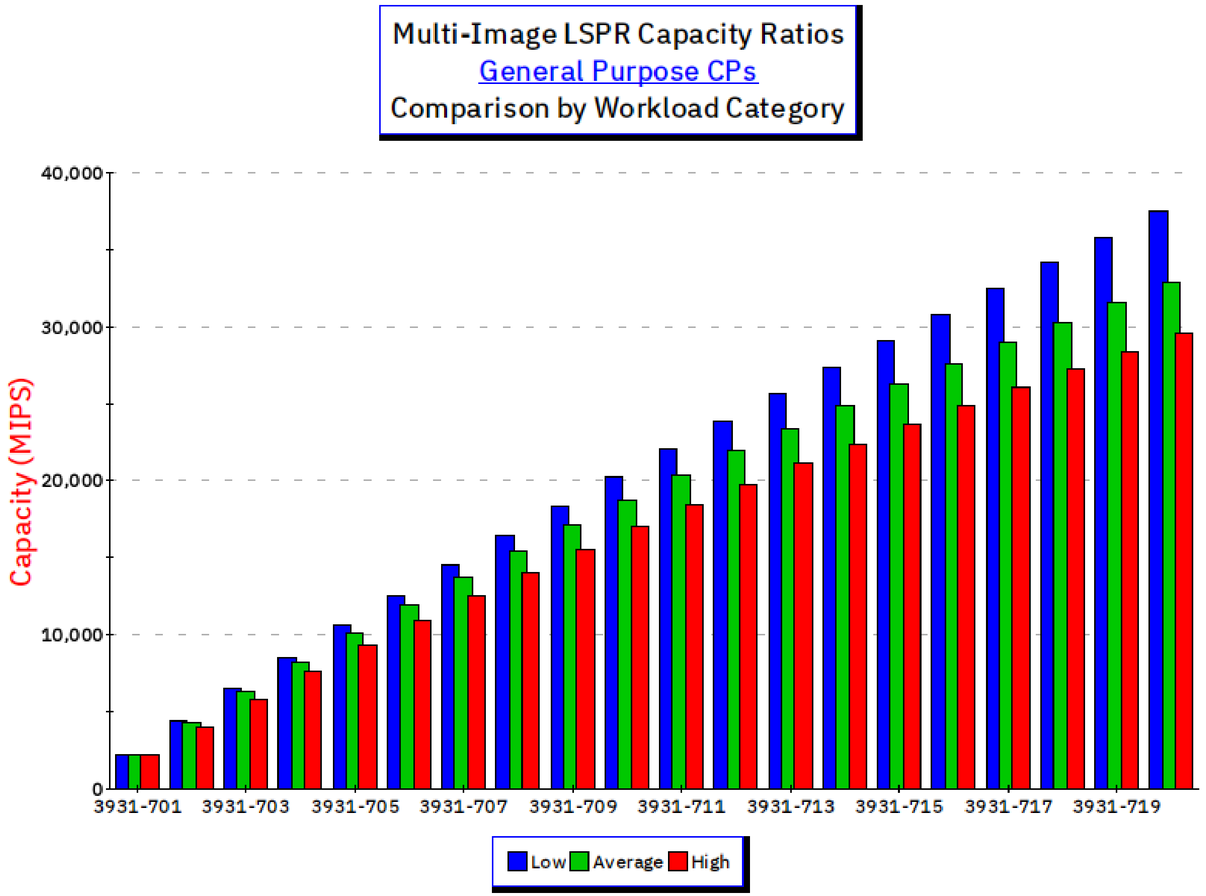
Accurately size IBM Z processors

Relative processor capacity varies by LPAR configuration and workload characteristics

- A single number MIPS table should not be used for detailed capacity sizing
- Use CPU MF data to match your workload to the appropriate LSPR workload curve on the ‘before’ machine to determine the optimal ‘after’ machine

Use zPCR which automatically processes CPU MF data to determine an LPARs LSPR workload match

- zPCR considers additional factors like system configuration, SMT benefit, and extends the workload categories to Low-Avg and Avg-High for unique cases



LSPR Processor Capacity Ratios

Value of CPU Measurement Facility (CPU MF)

Accurately size IBM Z processors

Washington Systems Center

Ex) z17 upgrade

- Standard upgrade, no growth or other factors considered
- Only using LSPR Workload Category – zPCR should be used to consider other factors

LSPR Match on z15	z15 715	z17 Processor Needed
Low	26,630	713
Average	23,632	712
High	20,841	711 or 712

LSPR Match on z16	z16 715	z17 Processor Needed
Low	29,122	714
Average	26,251	713 or 714
High	23,659	713

LSPR Multi-Image Capacity Ratios

Favorite CPs

Values are applicable for z/OS; representative of z/VM, KVM, and Linux

Capacity basis: 2094-701 @ 559.792 MIPS for a typical multi-partition configuration

Capacity for z/OS on z10 and later processors is represented with HiperDispatch turned ON

IBM Z Processor	Features	Flag	MSU	LSPR Workload Category				
				Low	Low-Avg	Average	Avg-High	High
8561-715	15W	=	2,819	26,630	25,042	23,632	22,149	20,841
3931-715	15W	=	3,108	29,122	27,612	26,251	24,887	23,659
9175-711	11W	=	2,702	24,061	23,309	22,602	21,640	20,757
9175-712	12W	=	2,894	26,050	25,154	24,317	23,272	22,312
9175-713	13W	=	3,079	28,011	26,959	25,984	24,857	23,824
9175-714	14W	=	3,256	29,944	28,726	27,604	26,398	25,294

Value of CPU Measurement Facility (CPU MF)

Validate achieved IBM Z processor performance

With CPU MF data on the "before" and "after" machines, you can use various levels of metrics to determine achieved performance relative to expectations

- Setting the expectation requires accurate LSPR workload category on "before" box
- CPI and LPAR CPU metrics required for high level comparison

See "How to Measure the New IBM z17"
Kwasi Abrefa-Kodom
Tuesday, May 13th at 2pm

High Level Metrics

Metric	Definition	Source	Level	Average ITRR	Median ITRR	Weighted
CPI	Cycles Per Instruction (GCP + zIIP)	SMF 113	LPAR	✓	✓	After LPAR CPU
IRATE / GCP	LPAR IO Interrupt RATE / LPAR GCP *	SMF 70 SMF 113	LPAR	✓	✓	After LPAR CPU
SSCH / GCP	DASD Start Subchannels / GCP	SMF72 SMF 113	LPAR	✓	✓	After LPAR CPU
LPAR Utilization	Utilization of the LPAR	SMF 70	LPAR			

* IRATE includes CTC and PCIE interrupts

Value of CPU Measurement Facility (CPU MF)

Gain efficiency insights: HiperDispatch and LPAR Configuration

IBM WSC Best Practice for defining logical CPs

– Define 0-2 more logicals than needed to meet CPs by weight
  - Should be the minimum of 25% additional capacity (small LPARs) up to 2 additional engines

– Don’t define all the logicals on a CEC to the LPAR

– Optimize for VHs

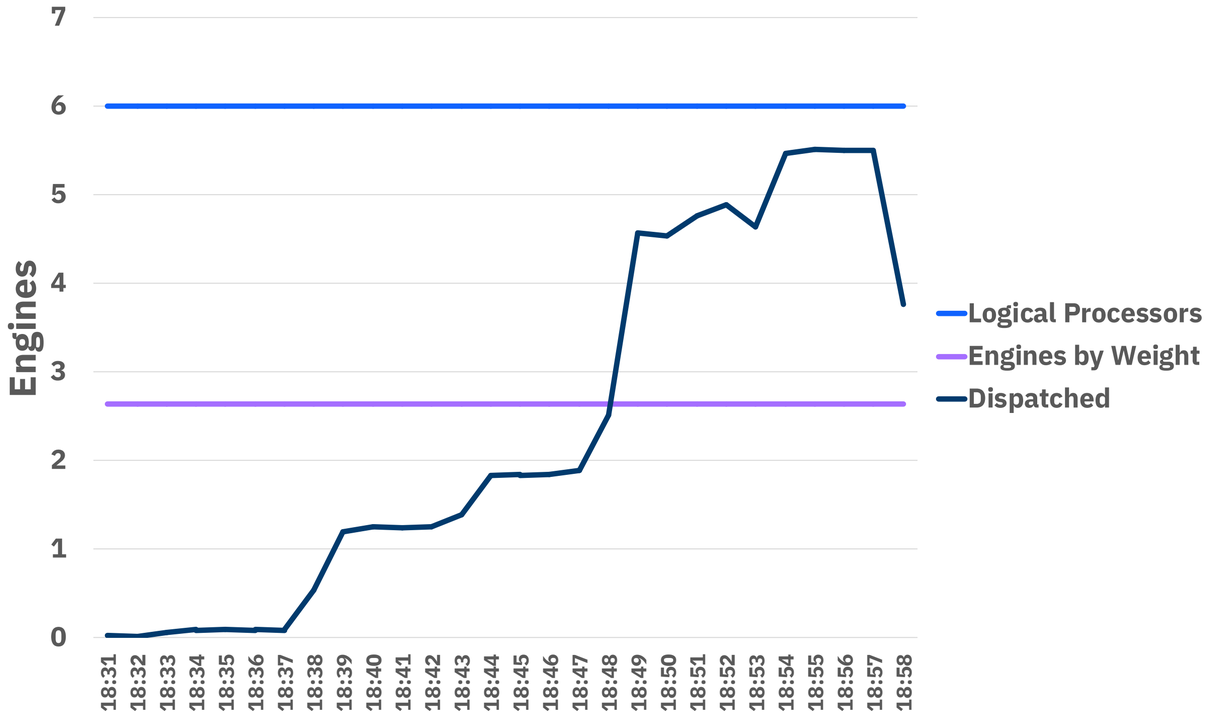
Why?

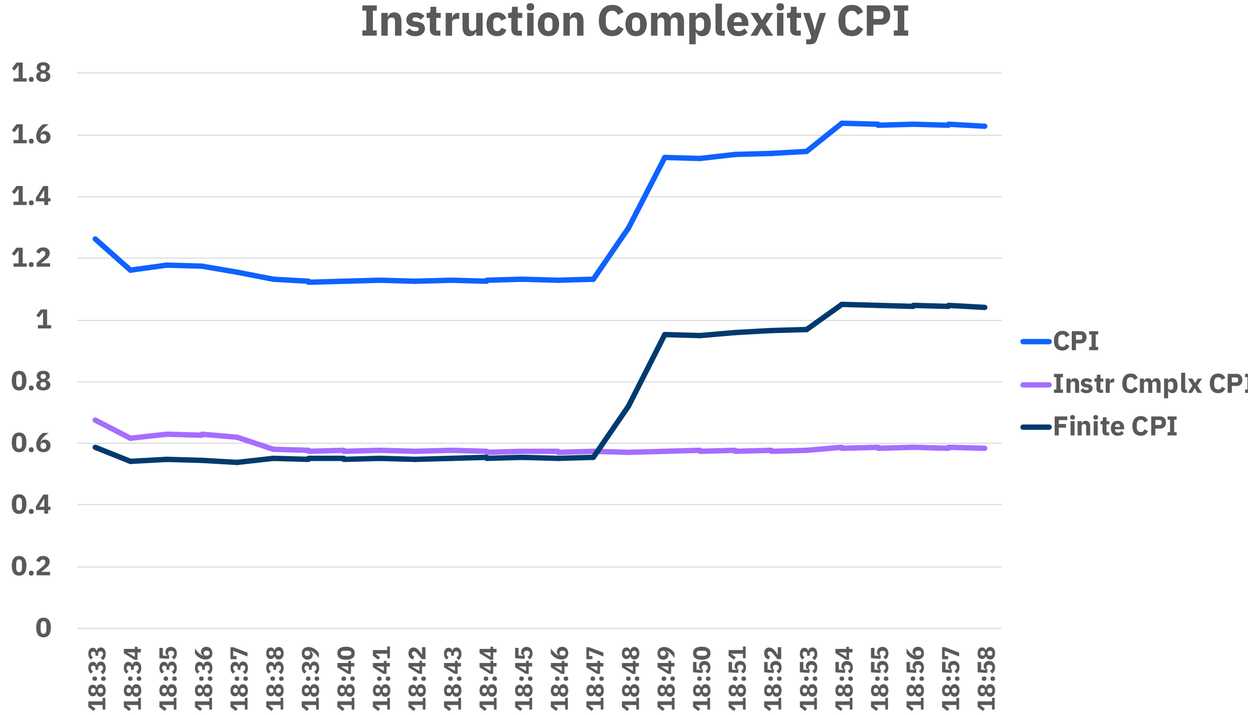
– Work runs most efficiently if you run with defined weight using VHs and VMs (can validate with CPU MF)
– LPAR Busy value displayed on online monitors is relative to number of LCPs
– LPAR time slice is sensitive to number of logicals, fewer logicals will lead to longer time slices
– Reduce the impact of a CPU Loop, fewer logicals limits potential impact
– Certain z/OS operations need to be done even for parked logicals
– Additional system resources utilized for each logical processor

Best practice whitepaper:
https://www.ibm.com/support/pages/node/6354843

Value of CPU Measurement Facility (CPU MF)

Gain efficiency insights: HiperDispatch and LPAR Configuration





– Running above weight on vertical lows led to increased CPI, which means it ran **less efficiently and *cost more***
– Remember, your vertical low is someone else's vertical high

Value of CPU Measurement Facility (CPU MF)

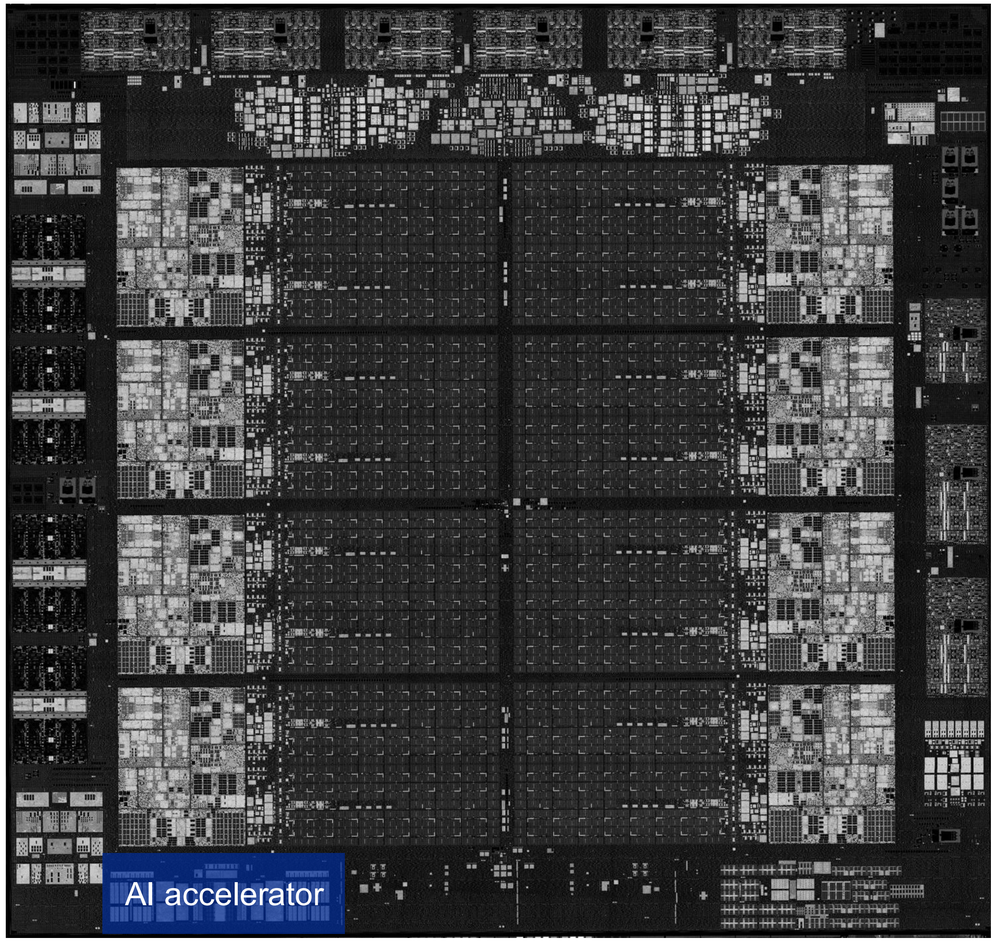
Gain efficiency insights: Measure AIU activity

IBM Washington Systems Center

New CPU MF metrics were introduced with IBM z16 Integrated Accelerator for AI

Centralized accelerator shared by all cores on-chip

- Very low and consistent inference latency
- Compute capacity for utilization at scale
- Designed for a variety of AI models ranging from traditional ML to RNNs and CNNs
- Security – provide enterprise-grade memory virtualization and protection



More AI acceleration on IBM z17

In-transaction AI with encoder LLMs and multiple AI model techniques

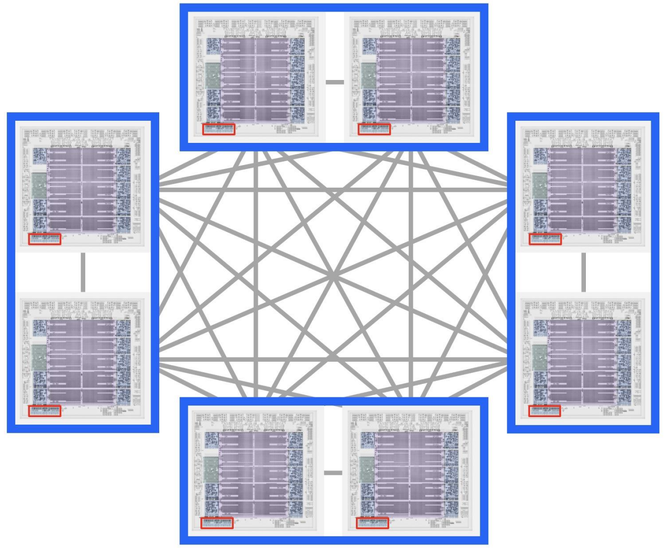
2nd Gen on-chip AI accelerator in Telum II

- Support for LLM compute primitives
- Improved quantization and matrix operations
- Improved AI processing over IBM z16[9]

AI workload balancing during peak usage

In-drawer intelligent routing

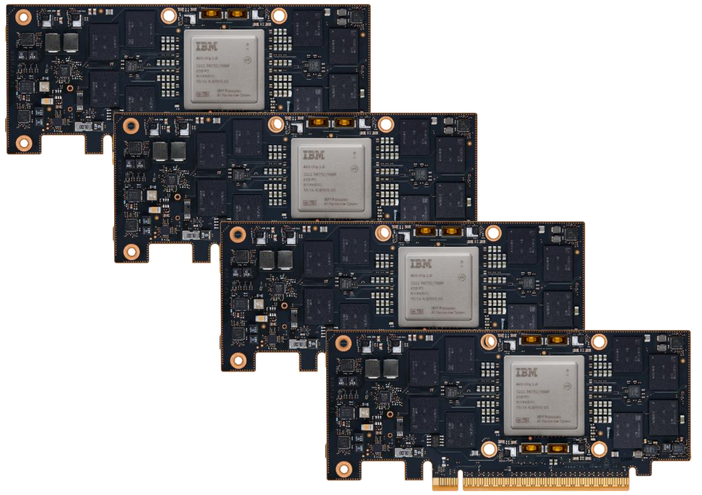
- Remote AI processing
- Up to 8x AI processing available



Optimize generative AI and LLM use cases

IBM Spyre Accelerator cards*

- 32 Gen AI-ready cores per adapter card
- Up to 48 adapter cards per system



*Available starting in 4Q25

Value of CPU Measurement Facility (CPU MF)

Gain efficiency insights: Measure AIU activity

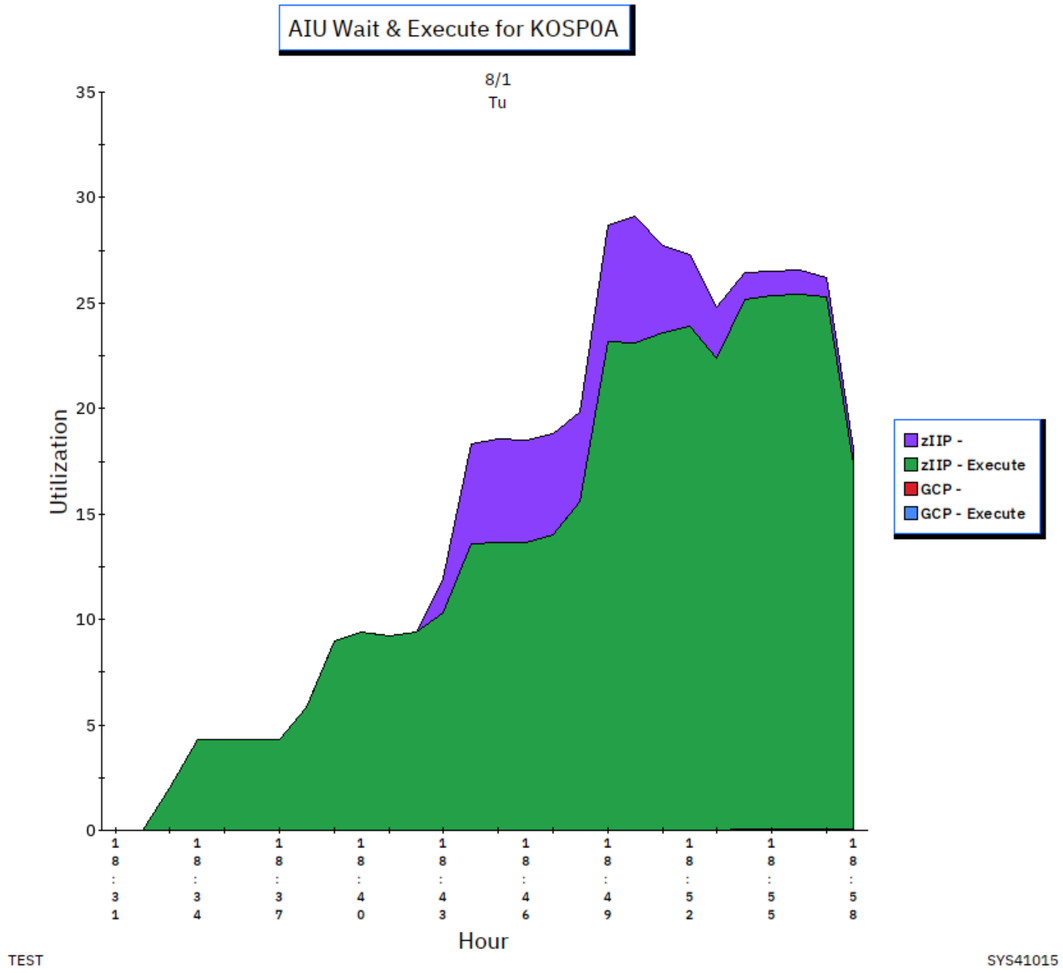
Existing metrics to understand load and demand for the Integrated Accelerator for AI on IBM z16 and z17

Executing on AIU % (C_AIU_CPU) - **Identify 'how much' AIU load**

- Need to consider how many AI accelerators the LPAR has access to
  - z16: One accelerator per chip, each core only has access to the accelerator on its chip
    - Ex. 8 cores on the same chip using the AIU = maximum value of 100% or 1 AI accelerator fully utilized
  - z17: One accelerator per chip, shared among all cores in the same drawer

Waiting on AIU % (W_AIU_CPU) - **Understand contention for the AIU**

- Time that a processor on a chip waits for another processor to finish with the AIU before it may execute on the AIU
  - Includes time spent obtaining the 'lock' for the AIU, indicating there will always be some 'wait' time



Value of CPU Measurement Facility (CPU MF)

Gain efficiency insights: Measure AIU activity

Washington Systems Center

New metrics to understand load, demand, and performance of the Integrated Accelerator for AI on IBM z16 and z17

<u>On average, how much time per request was spent executing vs waiting:</u>

Executing on AIU Time (C_AIU_Time)

– AVG microseconds spent executing on AIU per completed NNPA instruction

Waiting on AIU Time (W_AIU_Time)

– AVG microseconds spent waiting on AIU per completed NNPA instruction

z17 Only

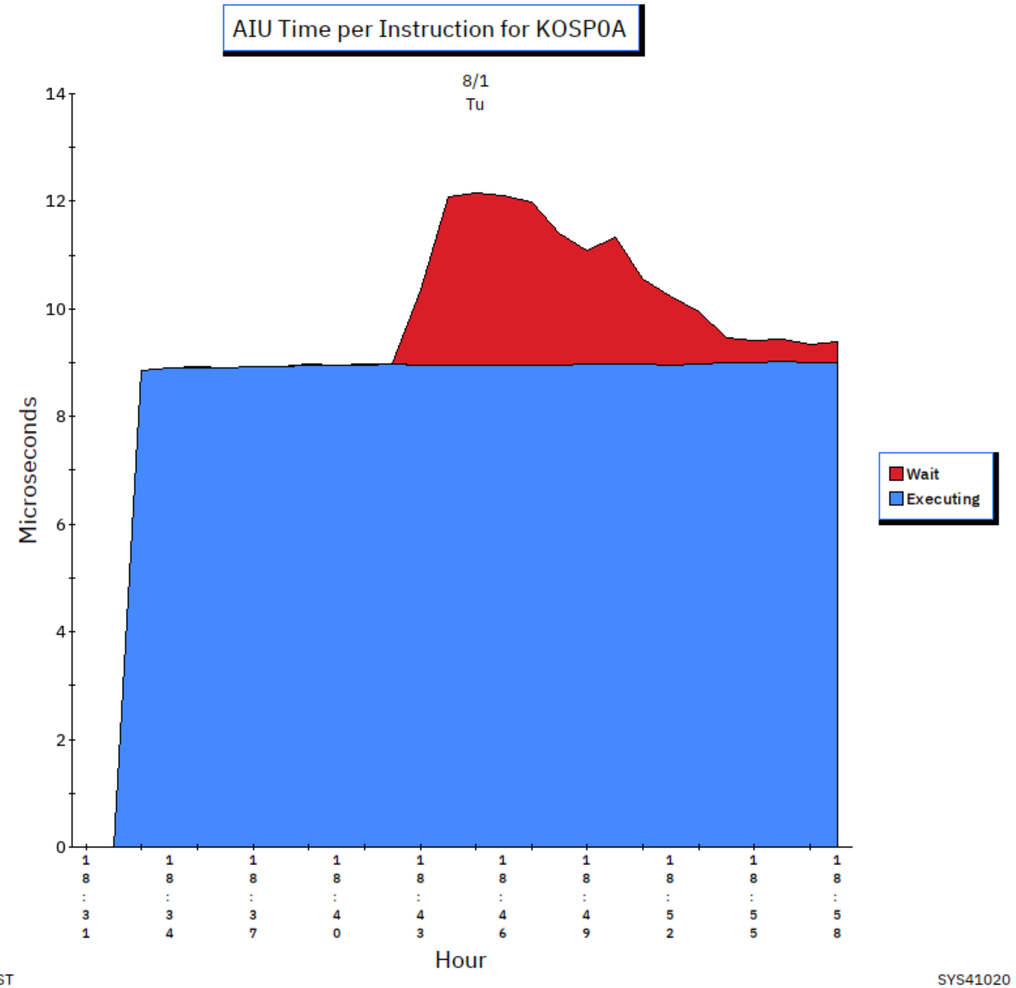
<u>What percent of NNPA requests were executed on local AIU vs remote:</u>

Local AIU %

– Percent of total NNPA instructions that were executed on the local, on-chip AIU

Remote AIU %

– Percent of total NNPA instructions that were executed on the remote, off-chip AIU



Value of CPU Measurement Facility (CPU MF)

Gain efficiency insights: Measure AIU activity

Washington Systems Center

AIU Metric Summary

	CPU MF Metric	Description	
	AIU_CPU	Total AIU CPU (LPARCPU Units: 100 = 1 ‘engine’)	
	C_AIU_CPU	Executing on AIU (LPARCPU Units)	
	W_AIU_CPU	Waiting for access to AIU (LPARCPU Units)	
NEW	**C_AIU_Time**	AVG microseconds spent executing per completed instruction	
NEW	**W_AIU_Time**	AVG microseconds spent waiting per completed instruction	
NEW	**LOCAL_AIU %**	Percent of NNPA invocations executing on local AIU	**z17 Only**
NEW	**REMOTE_AIU %**	Percent of NNPA invocations executing on remote AIU	**z17 Only**

Value of CPU Measurement Facility (CPU MF)

Gain efficiency insights: Identifying Store Into Instruction Stream (SIIS) Activity

CPU MF can be used to help identify when inefficient “SIIS” programs may be running

–Based on % of certain I Writes / D Writes sourced

–LPAR view, identifies when it happens, not who is causing it

- Look for repeating, high impact timeframes
- Identify the program(s) running in the time period, e.g. via zBNA Top Programs
- Use a hot spot analyzer to find the issue
- Remediate the source code to correct the issue

z17 formula is the same as z16

Processor	SIIS Indicator %	Description
z13	E163 / B2 *100%	I Writes sourced with L3 intervention / I Writes
z14	E163 / B2 *100%	I Writes sourced with L3 intervention / I Writes
z15	E164 / B2 *100%	I Writes sourced with L3 intervention / I Writes
z16	E170 / B2 *100%	I Writes sourced with L2 intervention / I Writes
z17	E170 / B2 *100%	I Writes sourced with L2 intervention / I Writes

Value of CPU Measurement Facility (CPU MF)

Gain efficiency insights: Identifying Store Into Instruction Stream (SIIS) Activity

IBM Washington Systems Center

Based on the SIIS Indicator %, the following actions are recommended

SIIS Description	SIIS Indicator %	Action
Noise – it will never be 0%	< 2%	None
Minimal SIIS impact	2% < 5%	Low Priority but potential MSU savings
Noteworthy SIIS impact	5% < 10%	Medium Priority – Investigate and Remediate
Considerable SIIS impact	>= 10%	Top Priority – Investigate and Remediate

With Tailored Fit Pricing, all MIPS count

More information

2006 TechDoc: IBM System z and eserver zSeries Processor Performance: Processor Design Considerations
– https://www.ibm.com/support/pages/system/files/inline-files/istream_flash_062606_v4.pdf

Dec 2019 TechDoc: Identifying “Store Into Instruction Stream” (SIIS) Inefficiency by Using CPU MF Counters
– Includes SIIS Assembler remediation examples
– https://www.ibm.com/support/pages/identifying-%E2%80%9Cstore-instruction-stream%E2%80%9D-siis-inefficiency-using-cpu-mf-counters

Value of CPU Measurement Facility (CPU MF)

Help with performance analysis: Cross-drawer impact

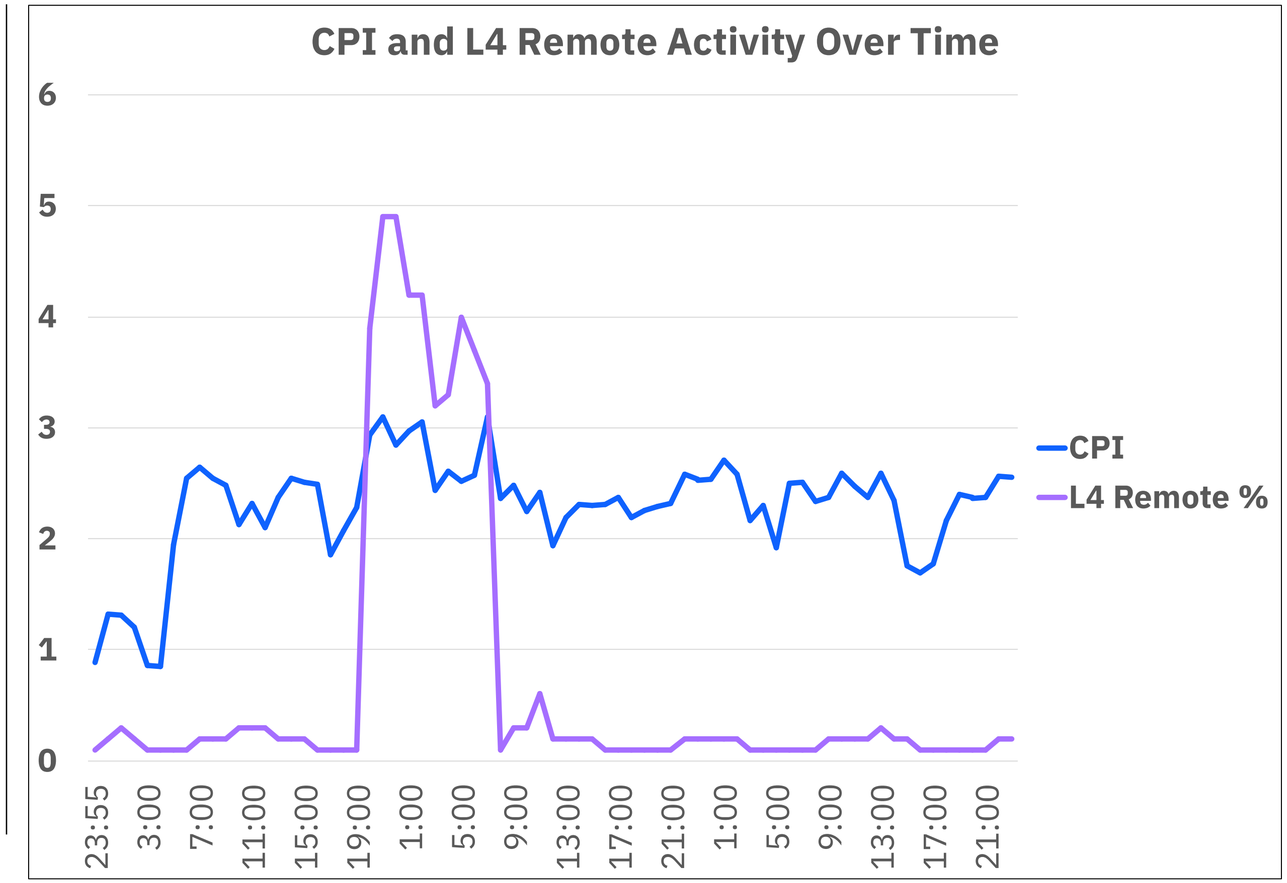
Understand the “why” behind performance issues

Ex) Cross-drawer impact

Performance analysis with CPU MF metrics
– Is there cross-drawer impact?
  • L4RP: Sourcing from remote (cross-drawer) L4

– Cost of the impact?
  • ESCPL1M: Avg sourcing cycles required per L1 miss
  • CPI: Cycles per instruction

Solution for some causes
– Large partition spanning drawers? Split LPAR (use zPCR to model and see expected capacity *gain*)
– Caused by VL usage? Optimize for VHs



Implementation: Enabling and collecting CPU MF counters

z/OS LPARs

– z/OS uses Hardware Instrumentation Services (HIS)

- *z/OS started task*
- *Writes SMF 113 records*
- *Implementation instructions: https://ibm.biz/cpumf-zos*
  1. Configure the LPAR to collect CPU MF via HMC
  2. Configure HIS and add SMF 113s to the appropriate statements in SMFPRMxx member
  3. Start collection of CPU MF data via START and MODIFY HIS commands
     - » F HIS,B,TT='runid',PATH='/HIS/',CTRONLY,**CTR=(ALL)**,SI=SYNC
     - » CTR=ALL -> Basic, Problem, Crypto, and Extended

z/VM LPARs

– z/VM uses Monitor Records

- *Writes Domain 5 (Processor) Record 13 (CPU MF Counters) records*
- *Implementation instructions: https://ibm.biz/cpumf-zvm*

Minimal overhead for collecting data

IBM Washington Systems Center

Summary

Enable and collect CPU MF counter data if not already

Ensure any capacity planning and performance tools are using the latest formulas

Use zPCR (or zCP3000) to properly size for z17 upgrade

Use CPU MF metrics to optimize for performance and efficiency

Claims/Disclaimers

9. The IBM z17 Telum II processor is designed to seamlessly scale peak AI workloads as each core on the chip can access each of the 8 integrated accelerators for AI. By allowing routing of inference requests to any idle IBM Integrated Accelerators for AI within the same drawer, the IBM Integrated Accelerator for AI can increase inference throughput by up to 7.5x as compared to IBM z16.

   DISCLAIMER: Performance results are based on internal tests exploiting the IBM Integrated Accelerator for AI for inference operations on IBM z16 and z17. On IBM z17, each IBM Integrated Accelerator for AI allows any CPU within a drawer to direct AI inference request to any of the 8 idle AI accelerators on the same drawer. The tests involved running inference operations on 8 parallel threads with batch size of 1. Both IBM z16 and z17 were configured with 2 GCPs, 4 zIIPs with SMT and 256 GB memory on IBM z/OS V3R1 with IBM Z Deep Learning Compiler 4.3.0, using a synthetic credit card fraud detection model (https://github.com/IBM/ai-on-z-fraud-detection). Results may vary.

IBM Spyre™ Accelerator is
in tech preview until Q42025

Thank you