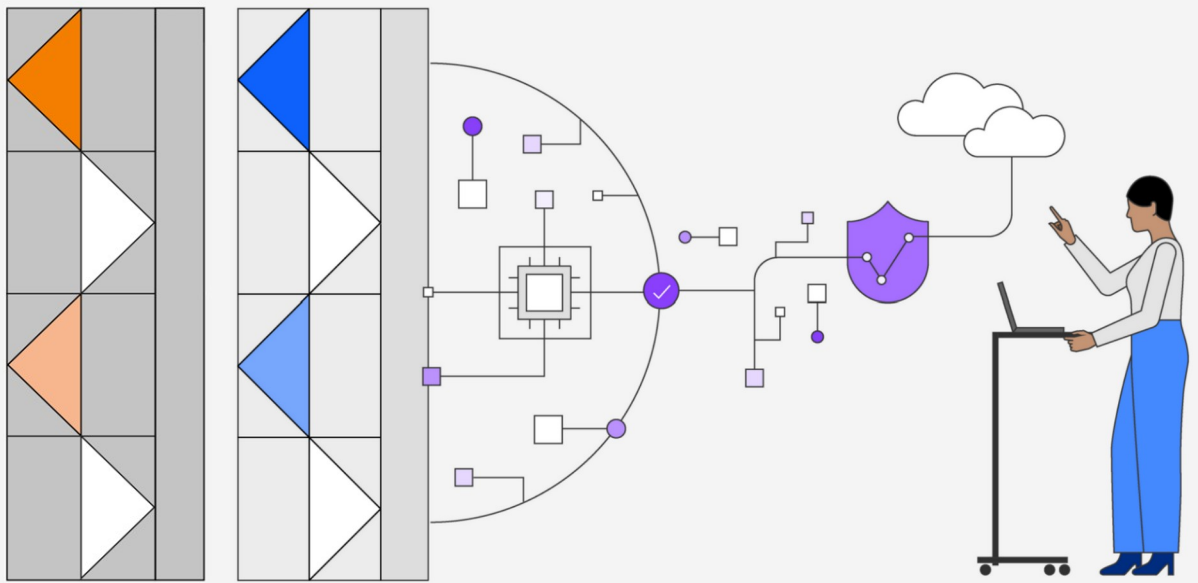
Put the Pedal to the Metal - Maximising Linux on IBM Z and LinuxONE Feature Usage

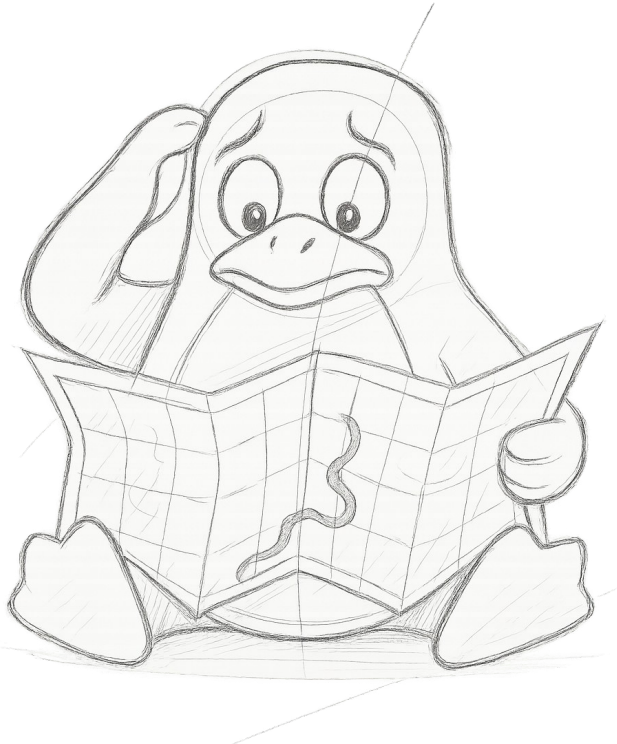
Stefan Raspl
Principal Product Manager
Linux & Virtualization on IBM Z and LinuxONE



Contents

So far, So good... Now what?!

So you got Linux on IBM Z or LinuxONE up and running, workloads are humming.
But are you really taking advantage of all the possibilities...?



So: Does it...work?!

Basics: CPU Counters Overview

- Ideally, accelerators work without add'l configuration or interaction
- But if they work transparently, how can you tell that they are in use...?
- (CPU) counters to the rescue! The two most popular ones on the platform:
  - **PAI**: Processor Activity Instrumentation (***preferred method***)
    - Introduced with z16
    - More fine-granular than CPUMF
    - Available in LPARs and VMs
    - See *Linux Device Drivers, Features, and Commands* chapter *Using the CPU Processor Activity Instrumentation Facility* for further details
  - **CPU-MF**: CPU Measurement Facility
    - Introduced with z10
    - Not virtualized, hence available in LPARs only, and therefore of limited use
    - See *Linux Device Drivers, Features, and Commands* chapter *Using the CPU-measurement facilities* for further details
- **Notes**:
  - Some tools like `cpacfstats` use both types of counters internally
  - Coverage of HW accelerators differs by framework
  - Needs enablement: All counters turned off by default due to minimal, but measureable, performance impact

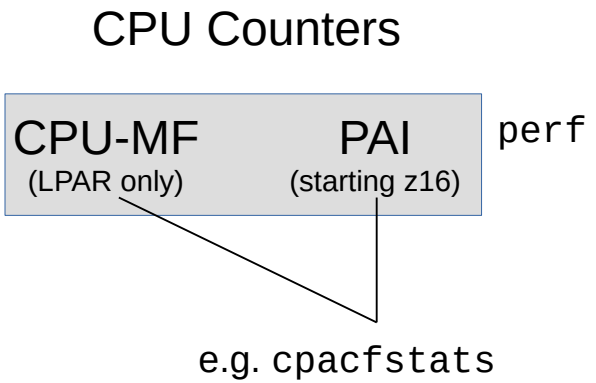
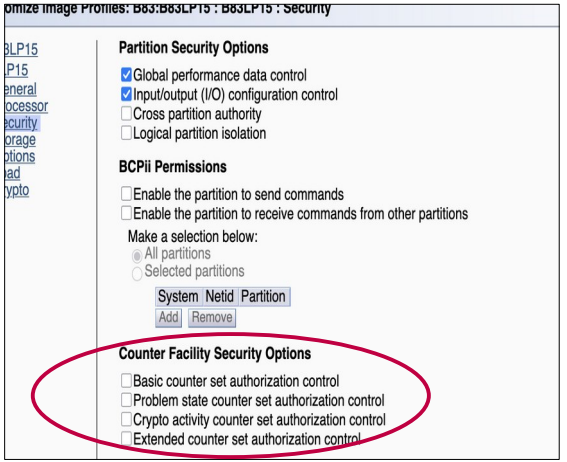


Fig. 1: Counters overview

Basics: CPU-MF

- Step 1: Enable appropriate counters in LPAR activation profile
  - Recommended: `Basic`, `Crypto` & `Extended`
  - Classic mode only: Don't forget to deactivate and activate the LPAR to make changes take effect
- Step 2: Verify general availability of counters in `lscpumf -i` command
- Step 3: Identify correct counter via `lscpumf -c` output
- Step 4: Collect data for a given counter.
  - Syntax: **`perf stat -e "cpum_cf/<counter>/" [-- <command>]`**
  - Specify `sleep <x>` as command to measure all instances in the entire Linux instance within the next x seconds




```
# Verify availability of counters
$ scpumf -i
CPU-measurement Counter Facility
------------------------------------------------------
Version: 3.8

Authorized counter sets:
    Basic counter Set
    Crypto-Activity counter Set
    Extended counter Set
    MT-diagnostic counter Set
    Problem-State counter Set

Linux perf event support: Yes (PMU: cpum_cf)
[...]

# Check available counters
$ lscpumf -c
perf event counter list for IBM z16
=====================================================
Raw event   Name    Description
-----------------------------------------------------
10:0    CPU_CYCLES
              Cycle Count
              Counter 0 / Basic Counter Set.
[...]

# Sample CPU cycles for 3 seconds
$ perf stat -e "cpum_cf/CPU_CYCLES/" -- sleep 3

        Performance counter stats for 'sleep 3':
         1,669,876      cpum_cf/CPU_CYCLES/

       3.000281025 seconds time elapsed

       0.000158000 seconds user
       0.000447000 seconds sys
```

Fig. 1: *CPU-MF fundamentals*

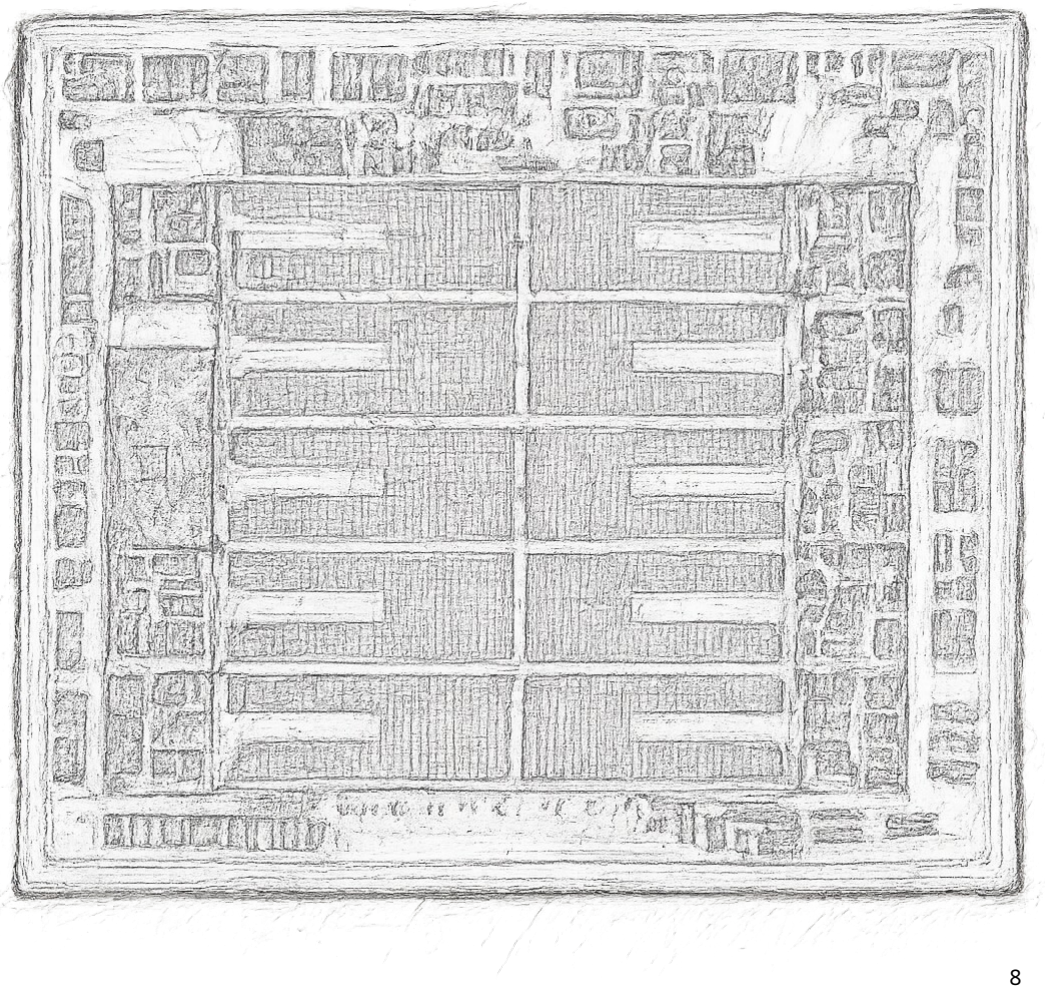
Basics: PAI

- Step 1: List available PAI counters using `lspai` to identify an appropriate counter
- Step 2: Collect data
  - `perf stat -a -e <counter(s)> -- [<workload>]`
  - Can collect per CPU, but for our purposes, aggregate via `-a` is just fine
  - See `/sys/devices/pai_crypto/events/` and Principles of Operation Chapter 5 for list of counters

```
# List all available counters
[root@a83lp73 ~]# lspai
RAW     NAME                 DESCRIPTION
12:4096 CRYPTO_ALL           Counter 0 / PAI CRYPTO counter set
12:4251 IBM_RESERVED_155     Counter 155 / PAI CRYPTO counter set
12:4252 IBM_RESERVED_156     Counter 156 / PAI CRYPTO counter set
[…]

# Collect sample data for all crypto counters
$ perf stat -a -e CRYPTO_ALL -- sleep 3

Performance counter stats for 'system wide':

              276      CRYPTO_ALL

      3.000731962 seconds time elapsed
```

Fig. 1: *PAI counter fundamentals*

On-Chip Features

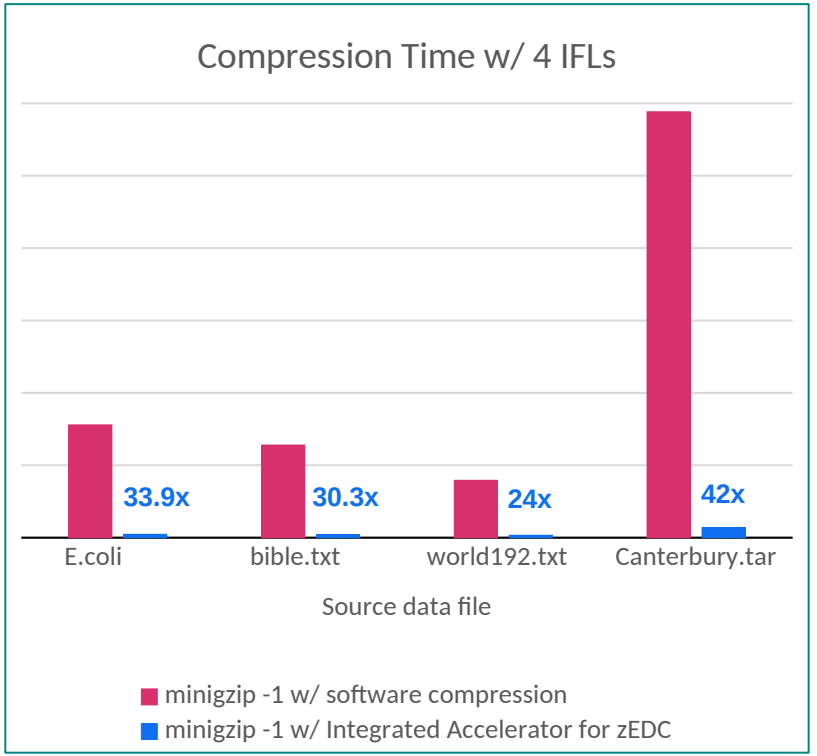


Deflate (aka NXU)

19.10 8.1 9 12 SP5 15 SP1 LPAR z/VM KVM

- **What it does**: Compress and uncompress data completely in hardware, providing massive performance increase
- Compression equivalent to `gzip -1`

  `-1` is fastest, `-9` slowest, default is `-6`
- Can be exploited e.g. by `zlib`, `gzip`, Java et al
- Compress data with zlib on IBM z15 with 4 processors up to 42x faster as compared to software compression
- **How to use**
  - Requires a z15 or LinuxONE III or later
  - Enabled for the **default compression level** starting with RHEL 8.4, SLES 15 SP3, and Ubuntu 20.10 and all later releases
    - Use env variable `DFLTCC_LEVEL_MASK` to enable for arbitrary compression levels
  - Java: Use Java 8 SR6 FP16 or later on any Linux distribution
  - See here for further details on usage



Deflate (aka NXU) - *continued*

- **How to verify**
  - Availability reported with new feature flag in `/proc/cpuinfo`: `dflt`
  - Use **CPU-MF** counter `DFLT_CYCLES`: Increases on use of NXU
    - Note: No PAI counter available!
  - Alternative: Confirm by runtime comparison
    - Userspace: Use environment variable `DFLTCC_LEVEL_MASK` to toggle use of feature
    - Kernel: Reboot with kernel parameter `dfltcc=off`

```
# Use perf to verify usage
$ perf stat -e "cpum_cf/DFLT_CYCLES/" -- gzip foo

Performance counter stats for 'gzip foo.txt':

      685,067,259      cpum_cf/DFLT_CYCLES/

      0.472829343 seconds time elapsed

      0.141366000 seconds user
      0.331596000 seconds sys

# Measure time with NXU turned off:
$ time DFLTCC_LEVEL_MASK=0x0000 gzip <file>

# Measure time with NXU turned on:
$ time DFLTCC_LEVEL_MASK=0x01ff gzip <file>
```

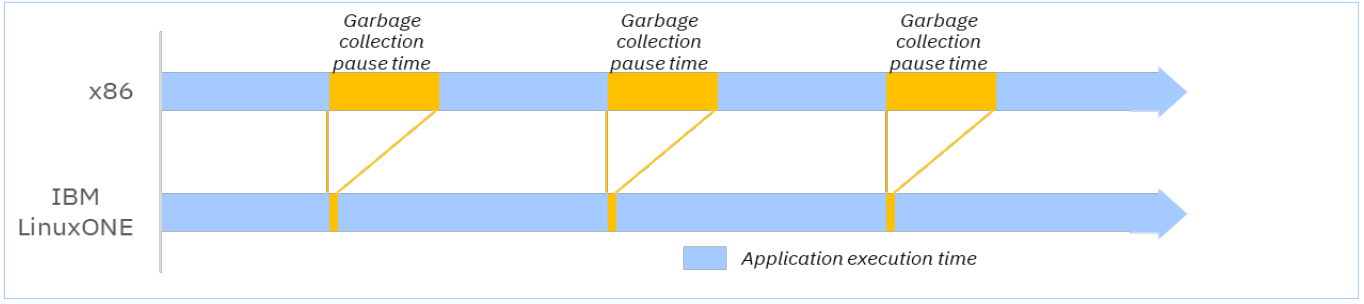
Fig. 1: *Sample verification of NXU usage*

Pause-Less Garbage Collection `-Xgc:concurrentScavenge`

LPAR z/VM KVM

- **What it does**
  - Situation ***before***: all applications threads are paused – hence the term *stop-the-world* operation – to allow GC to safely move objects in heap
  - ***Pause-Less GC***: short GC pause times, application threads run ***concurrently*** with GC
  - More *consistent* and *reduced* GC pause times for ***response time sensitive***, large heap applications
  - The "Concurrent Scavenge" GC allows garbage collection to run virtually concurrently with application threads without having to pause the JVM, hence providing greatly reduced GC pauses.[1]

- **How to use**
  - Prerequisites:
    - IBM z14 or LinuxONE II or later
    - Java V8 SR5
    - RHEL 7.5, SLES 12 SP4, and Ubuntu 17.10, or any later release
  - Specify the "`concurrentScavenge`" mode for garbage collection on the cammand line via "`-Xgc:concurrent Scavenge`"



[1] https://blog.openj9.org/2019/03/25/concurrent-scavenge-garbage-collection-policy/

Pause-Less GC: `-Xgc:concurrentScavenge`

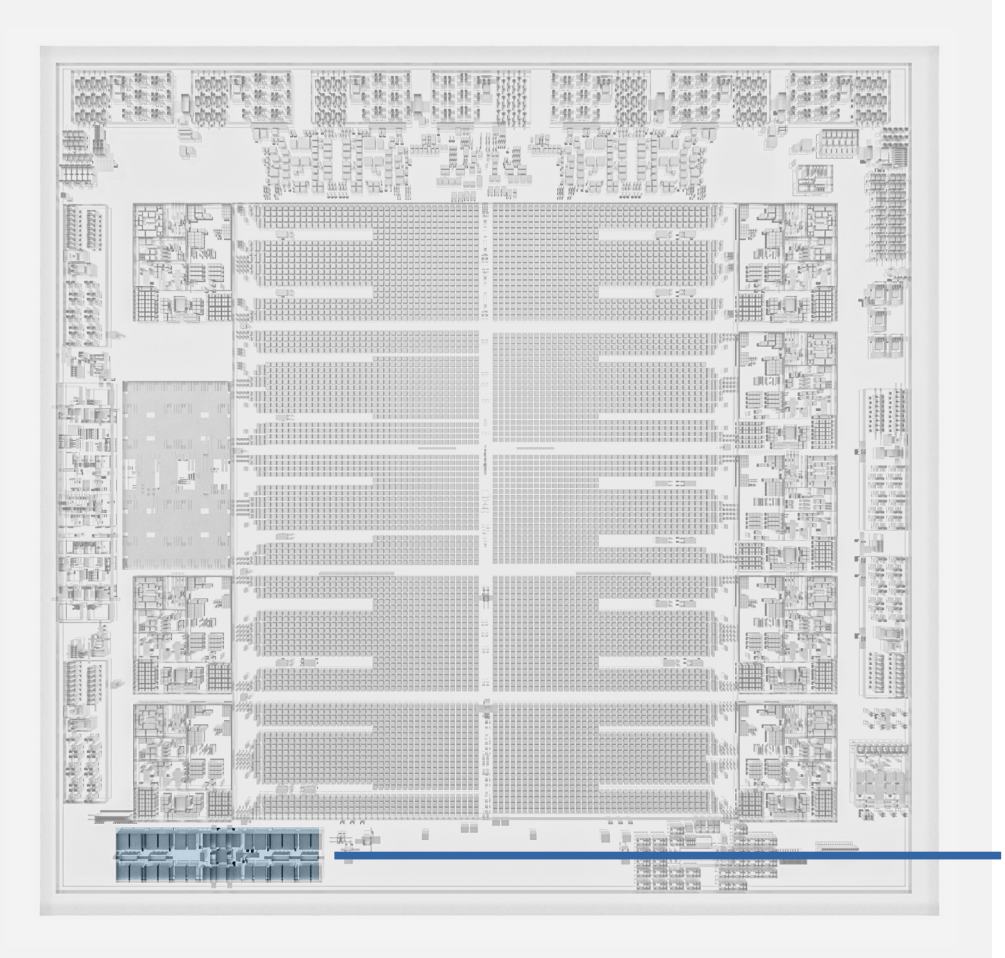
LPAR z/VM KVM

- **How to use**
  - Prerequisites:
    - IBM z14 or LinuxONE II or later
    - Java V8 SR5
    - RHEL 7.5, SLES 12 SP4, and Ubuntu 17.10, or any later release
  - Specify the *concurrentScavenge* mode for garbage collection on the cammand line via `-Xgc:concurrentScavenge`
- **How to verify**
  - Activate verbose garbage collection log, and verify that `concurrentScavanger` is enabled and indicates `H/W assistance`:

```
<?xml version="1.0" ?>
 <verbosegc xmlns="http://www.ibm.com/j9/verbosegc" version="48fc32a_CMPRSS">
 <initialized id="1" timestamp="2026-01-07T05:43:38.995">
 <attribute name="gcPolicy" value="-Xgcpolicy:gencon" />
 <attribute name="concurrentScavenger" value="enabled, with H/W assistance" />
```

Improved On-processor AI Acceleration

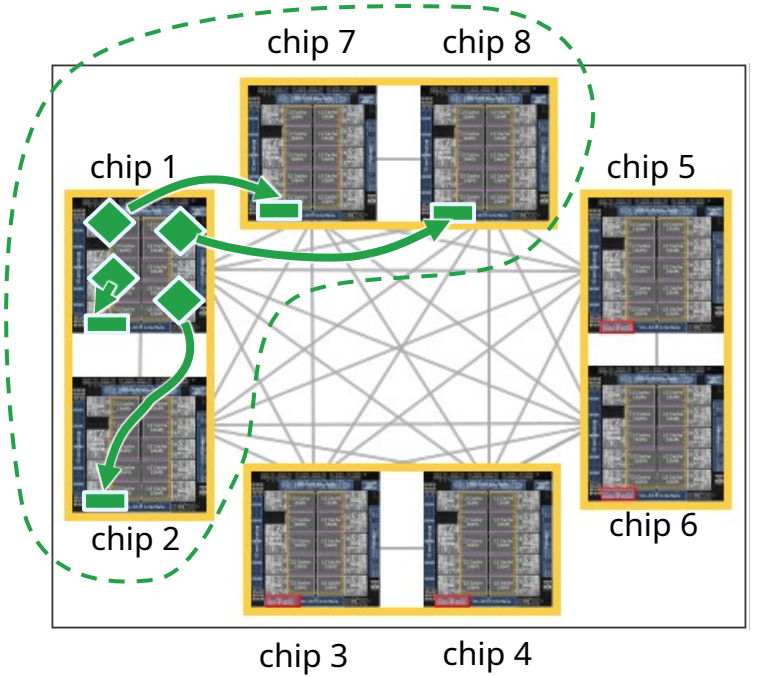
- Integrated as a CISC instruction
- Remote AI accelerator
- Support for LLM compute primitives
- Int8, FP16 datatypes



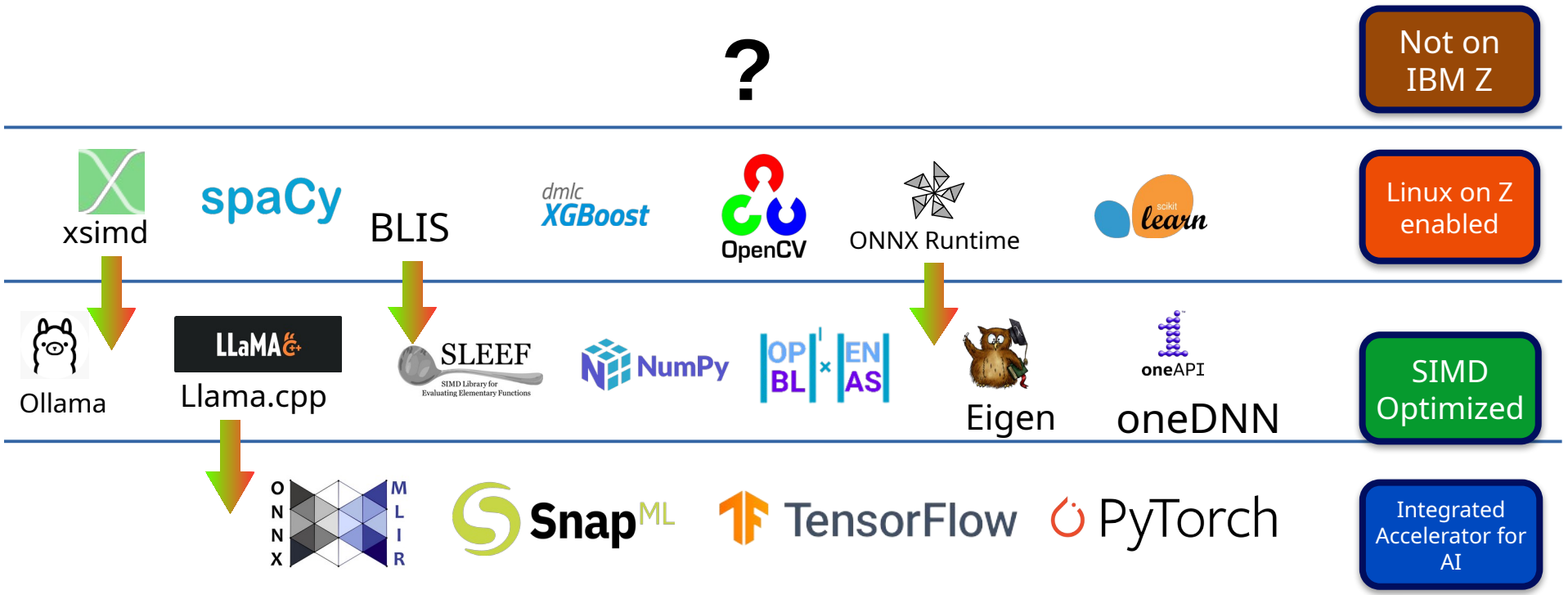
Telum II: Improved On-Chip AI Acceleration

- Integrated as a CISC instruction
- Support for LLM compute primitives
- Int8, FP16 datatypes
- Remote AI accelerator

- On z16, all cores on a chip (up to 6 user cores) would share the local on-chip accelerator; what can lead to serialization of AI work running on the same chip
- There are 8 on-chip accelerators available within a drawer
- z17 and LinuxONE 5: these accelerators are shared across all cores in that drawer, allowing transparent load balancing of AI work on all 8 accelerators
- AI work is spread across accelerators transparently always using the closest to the AI workload accelerator and giving on average equal runtime on the accelerator to all AI workloads
- Note: There is some penalty for using a remote accelerator



AI Ecosystem
Hardware Exploitation: How to use



On-chip AI

LPAR z/VM KVM

- **How to verify**
  - Use PAI counter NNPA_ALL
  - E.g. run

    `perf stat -a -e NNPA_ALL -- sleep 10`

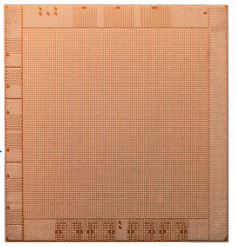
    while running a workload to catch all use of on-chip AI within the next 10 seconds

```
# Use perf to verify usage while
# workload is running
$ perf stat -a -e NNPA_ALL -- sleep 10

Performance counter stats for 'system wide':

               537      NNPA_ALL

     10.001517844 seconds time elapsed
```

Fig. 1: *Sample verification of NNPA usage*

Security

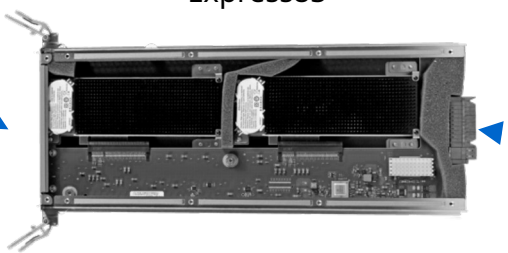
Overview – HW Crypto support in LinuxONE



PU SCM
Each PU is supports the CPACF function (clear key and protected key)

PCIe I/O drawers

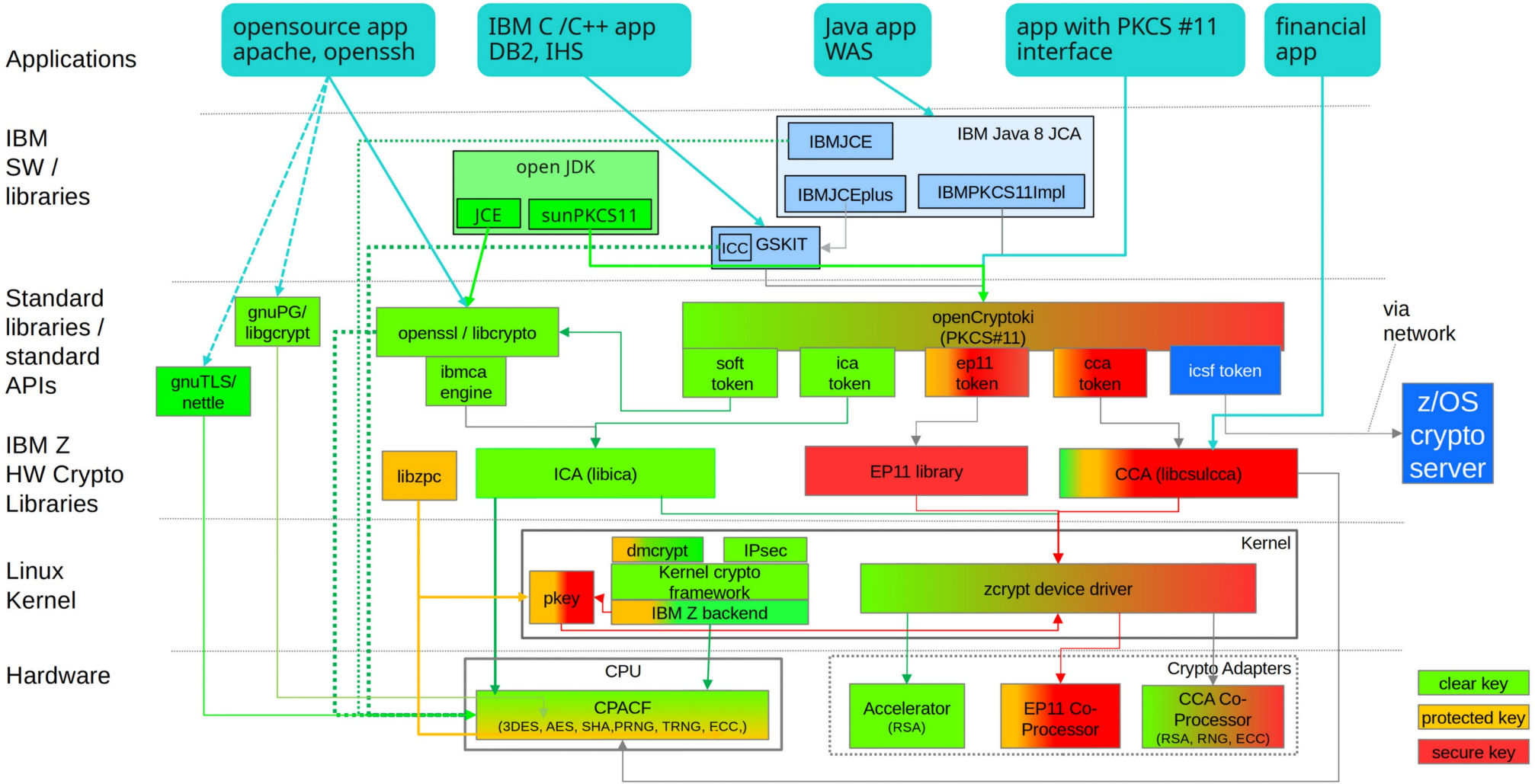
Crypto Express8S



Trusted Key Entry (TKE)

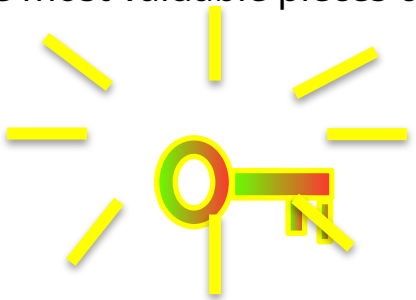
Smart Cards

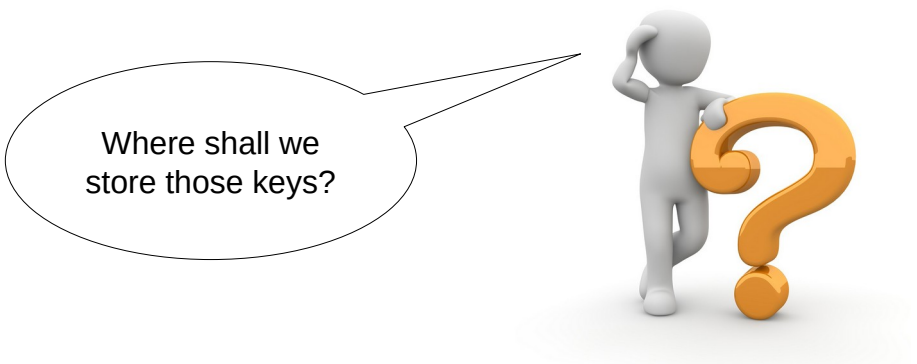
Protected Keys: How to Use



Protecting Encryption Keys: Catch 22?

Once our data is encrypted, our cryptographic keys become the most valuable pieces of data.





In a file?

- How then do we protect that file? That disk?
  - By encryption ???
    - ….

On a PC: enter the password / key at the console during boot

- Well, that may work for a laptop
- No good for a server (farm)!
- How to automate the server boot process?

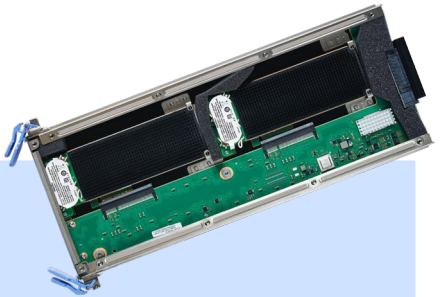
How do we protect keys from being stolen from main memory?

- How do we avoid creating dumps that contain keys?

Protecting keys by a Hardware Security Module (HSM)

Keys shall be protected by tamper responding HW then

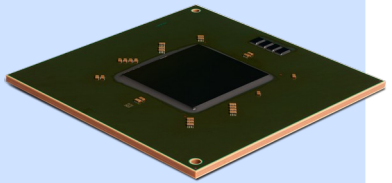
- Key values are never stored in plain text in the operating system memory
- Can only be used by someone who can access your physical HSM
- Can be stored securely on any (unprotected) media
- Cannot be stolen.



Combine

- IBM's HSM Technology:
  - **Crypto Express Adapters (CCA or EP11)**
    - » secure keys protected by a master key in a tamper responding device

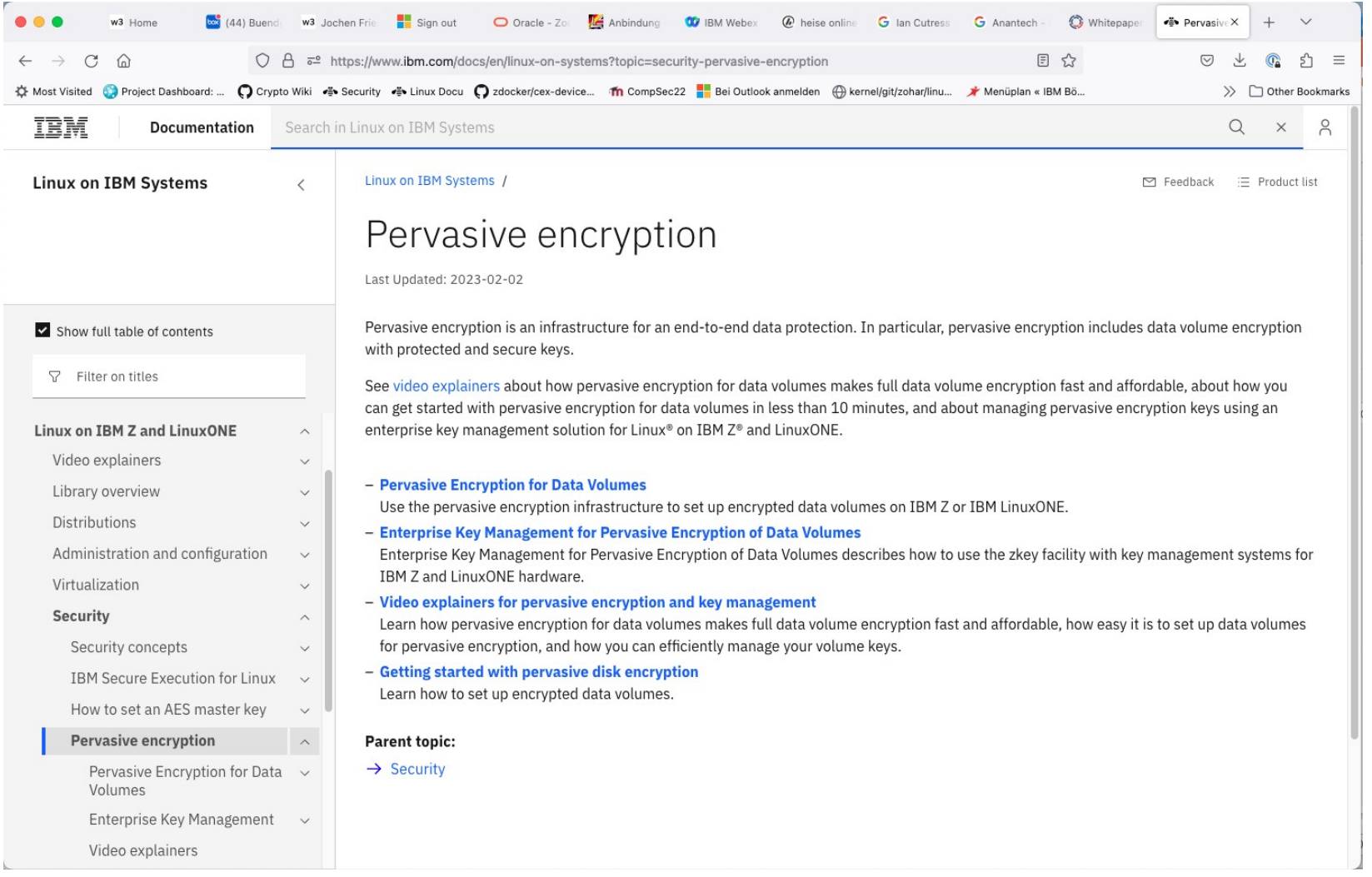
With



- IBM crypto acceleration:
  - **CPACF protected keys**
    - » Wrapped by a virtual server specific master key hidden in firmware
    - » Only valid while virtual server is running

Protected Keys: How to Use

Access documentation here



CPACF: How to verify

- PAI or CPU-MF counters: You know what to do...
- **Better: Use `cpacfstats` command**
  - Ships with `s390-tools` package
  - Needs to run `cpacfstatsd` daemon with root permissions
- User can access information provided by the daemon through the `cpacfstats` tool
  - Enable/disable CPACF Crypto Activity counters
  - Display counters after use (with optional filter `-n` to skip zero counters) to show usage of the CPACF functions
  - Reset counters to 0 via `-r`
  - Convert output to JSON format

```
# Start cpacfstats daemon before we go to work
$ cpacfstatsd

# Check current status of stats
$ cpacfstats
des counter: unsupported
aes counter: unsupported
sha counter: unsupported
rng counter: unsupported
ecc counter: unsupported
pai_user   : disabled
pai_kernel : disabled

# Enable statistics
$ cpacfstats -e -n
des counter: unsupported
aes counter: unsupported
sha counter: unsupported
rng counter: unsupported
ecc counter: unsupported
pai_user   : enabled
 ( 32) KMA AES 256bit                : 1537
pai_kernel : enabled
 (128) PRNO TRNG                     : 1

# Run specific workload and check again
$ openssl speed -hmac SHA-256
[...]
```

CPU-MF counters!

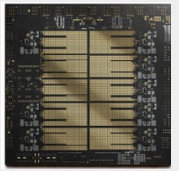
Fig. 1: *Preparing for cpacfstats use*

```
$ cpacfstats -n
des counter: unsupported
aes counter: unsupported
sha counter: unsupported
rng counter: unsupported
ecc counter: unsupported
pai_user   : enabled
 (  7) KM AES 128bit                 : 28
 (  9) KM AES 256bit                 : 692
 ( 32) KMA AES 256bit                : 3819
 ( 72) KIMD SHA1                     : 97
 ( 73) KIMD SHA256                   : 522
 ( 74) KIMD SHA512                   : 40
 ( 76) KIMD SHA3-256                 : 5
 ( 81) KIMD GHASH                    : 44
 ( 82) KLMD SHA1                     : 28
 ( 83) KLMD SHA256                   : 6
 ( 86) KLMD SHA3-256                 : 1
 ( 87) KLMD SHA3-384                 : 2
 ( 88) KLMD SHA3-512                 : 1
 ( 89) KLMD SHAKE 128                : 27
 ( 90) KLMD SHAKE 256                : 6
 (163) KMAC HMAC SHA 384             : 25
 (124) PCC Scalar Mult X25519        : 3
pai_kernel : enabled
 (128) PRNO TRNG                     : 1769

# Don't forget to turn off after use!
$ cpacfstats -d
[...]
```

Fig. 2: *Sample cpacfstats usage*

Post Quantum Cryptography

- **What it does**: Protect against “harvest now, decrypt later” attacks
- **How to use**: Choose the right ciphers:
  - **Public Key**: ML-KEM (aka Kyber)
  - **Digital Signatures**:
    - ML-DSA (aka Dilithium)
    - SLH-DSA (backup)
  - **Symmetric Encryption**: AES-256
- **How to verify**: See `cpacfstats`



IBM z17 Telum II Chip

IBM z17 provides hardware integration of on-chip quantum-safe encryption supporting Quantum key distribution (QKD) and quantum-resistant cryptography (QRC).

- No add-on SW or HW
- No system overhead
- No management required
- EAL 5+ level security

- *IBM z17 quantum-safe secure boot technology helps to protect IBM zSystems firmware from quantum attacks through a build-in dual signature scheme with no changes required.*
- *IBM z17 quantum-safe technology and key management services, were developed to help you protect data and keys against a potential future quantum attack like harvest now, decrypt later.*
- *Quantum-safe encryption technologies used in IBM z17 are designed to help protect against unintended disclosure of data throughout its end-to-end life cycle*

Secure Execution for Protecting “Data in Use”

Your Journey to Confidential Computing

- **What it does**: Allows users to run their Linux workloads with maximum privacy by protecting system memory.
- **Why you should care**:
  - Not even system/KVM host administrators can access customer data in KVM guests
    ⇒ Protection against insider attacks
  - Allows users to run sensitive workloads on and off premise with the same level of data protection
  - Reduces the efforts of a cloud service provider to establish and document procedures for compliance and certification
- **How to verify**: Use `pvattest`

Digital Assets
NFTs (Web3)
Policy Platforms

Confidential AI (specialized and foundational) protects data and models

Secure Containerized Workloads on Open Platforms

Secure Execution: How to Verify

Validation Target Environment

```
# Create attestation request

$ pvattest create -k host_key_document.crt --crt CA.crt \
--crt IBM_signing.crt --arpk arp.key --output attestreq.bin

Successfully generated the request



# Execute request and store results in measurement.bin

$ pvattest perform attestreq.bin measurement.bin
```

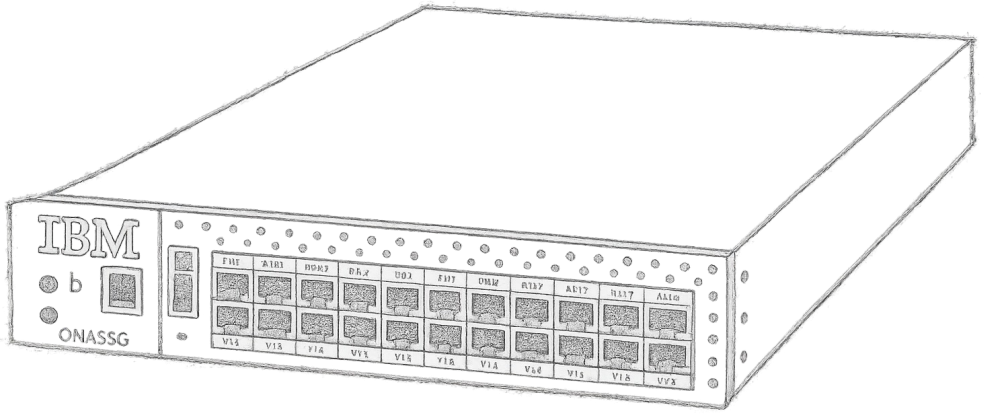
Fig. 1: *Prepare and execit attestation request*

Trusted Verification Environment

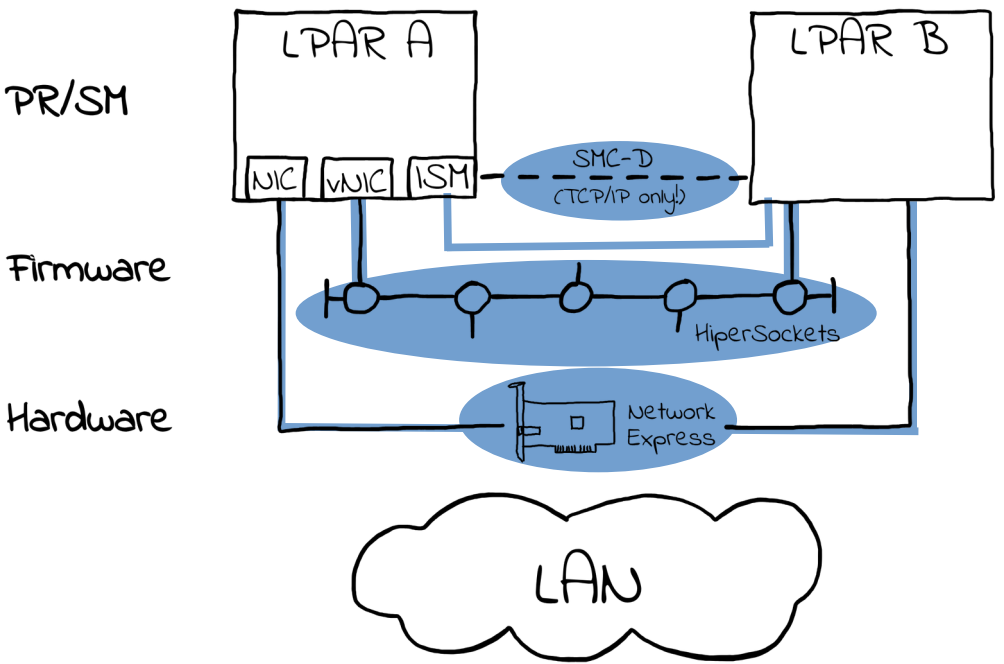
```
# After transferring ‘measurement.bin’, verify the results

$ pvattest verify --format=yaml --input measurement.bin \
--arpk arp.key --hdr secure_guest.hdr --verbose

Attestation measurement verified

      Config UID:

      0x14141414141414141414141414141414
```

Fig. 2: *Verify the results*

Networking

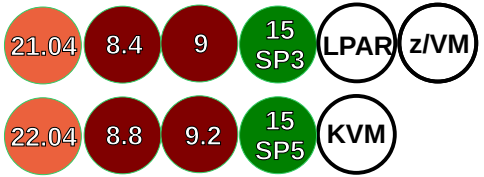


Virtual Networking Overview



- **What it does**: Multiple ways to provide fast communication between partitions
- **Why you should care**: Reduces latency, and, depending upon technology, can reduce CPU consumption and eliminate network cabling nightmare
- **How to verify**:
  - Use well-known tools like ip stats for HiperSockets and Network Express shared traffic
  - SMC-D: Read on…

Co-Location: SMC-Dv2



- **What it does**: Provides acceleration for TCP traffic
- **Why you should care**: v2 lifts limitations and greatly simplifies usage
- **Recap**
  - **Shared Memory Communications – Direct** provides intra-CEC acceleration for TCP traffic using *Internal Shared Memory* (ISM) devices
  - **Superior performance** (low latency, high throughput) at reduced CPU consumption
  - *However, SMC-Dv1 had limitations*:
    - Peers must be in **same IP subnet**
    - Devices need to be **paired using PNET IDs**
- **SMC-Dv2**
  - Peers can be in ***any* IP subnet**
  - No PNET IDs required
⇒ **Simplified configuration!**
  - Requires z15 or LinuxONE III
  - As with SMC-Dv1: Full **z/OS compatibility**
- **New performance paper available here**

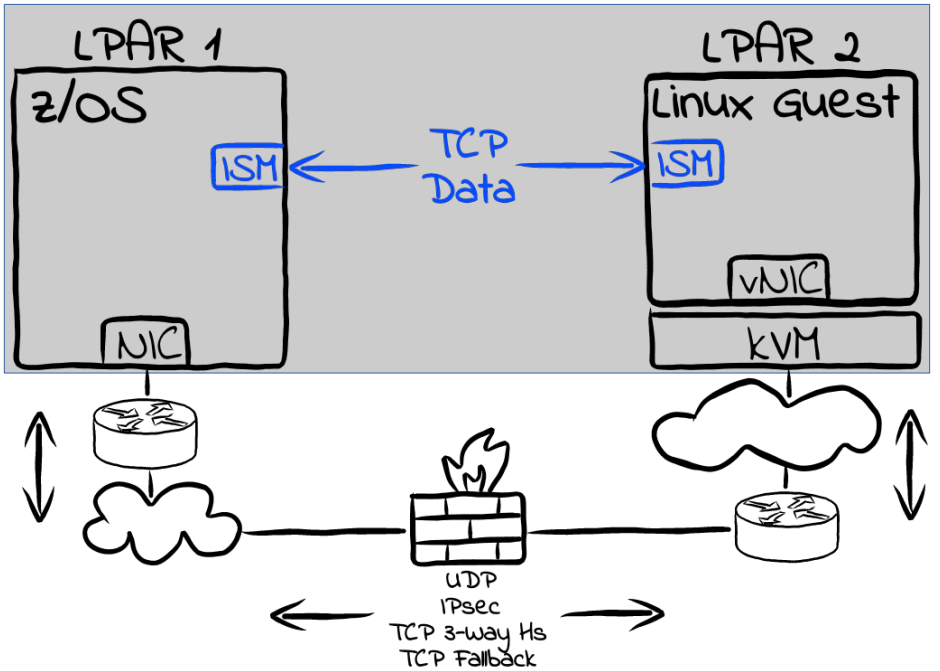


Fig.1: *Traffic flows with SMC-Dv2*

SMC-Dv2 – How to Use

- **IBM Z hardware requirements**
  - IBM z15 or LinuxONE III
  - DPM supported starting with IBM z16 / LinuxONE Emperor 4
  - Classic mode only
- ***Internal Shared Memory* (ISM) devices**
  - *Virtual* PCI network adapter of VCHID type ISM
  - 32 ISM VCHIDs per CPC, 255 FIDs per VCHID ⇒ 8K FIDs per CPC total)
  - I.e. maximum of 255 virtual servers communicating over same ISM VCHID
  - Each ISM device currently handling up to 1,920 connections
  - Assign multiple ISM devices to increase connection limit

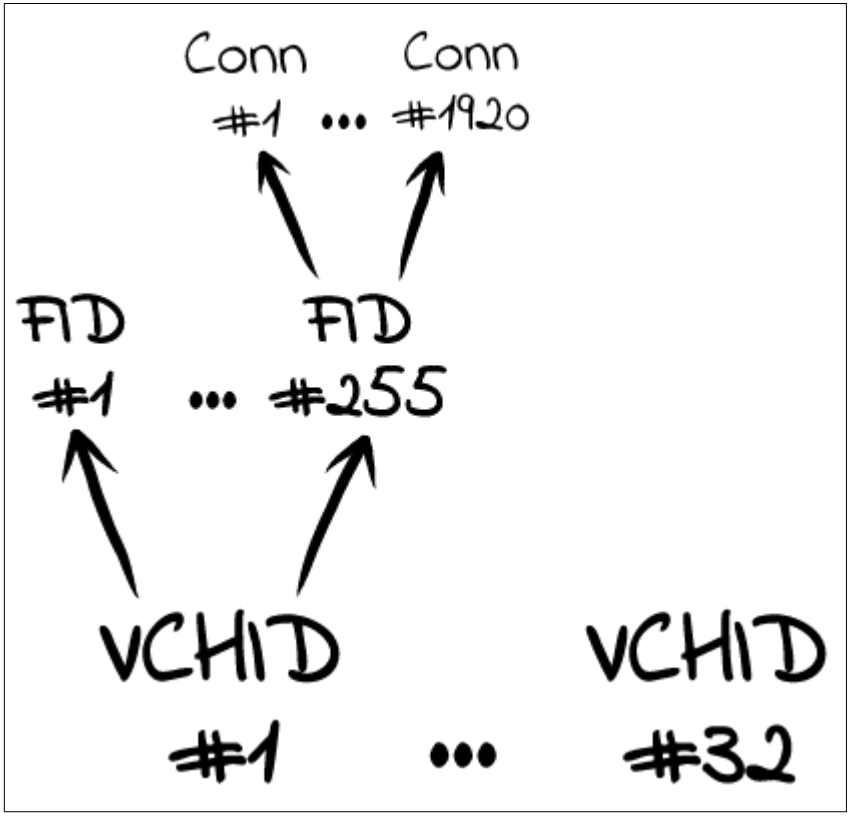


Fig.1: *Relationship between VCHIDs, FIDs and connections*

SMC-Dv2 – How to Use

- **SMC-Dv2 ISM device eligibility:**
  - (***recommended***) ISM devices without PNET ID
  - ISM devices with PNET ID matched by any networking interface (SMC-Dv1 compatibility)

- **ISM Device Setup**
  - Assign an ISM Device *without* a PNET ID
  - **`smc_rnics`**: Hotplug ISM devices, verify ISM presence, and check PNET IDs

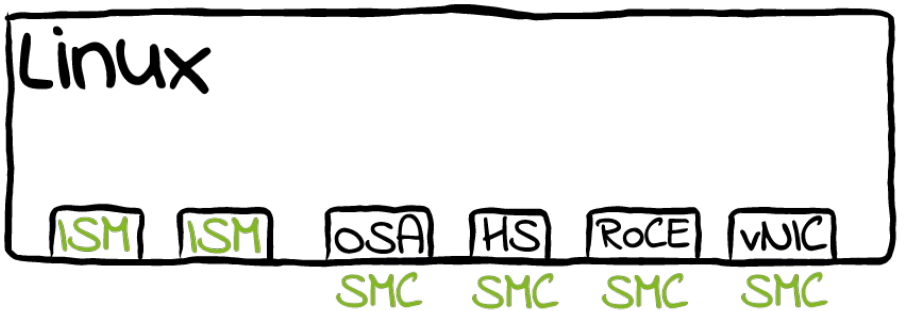


Fig.1: **Any** *interface is enabled for SMC-Dv2 – no further per-interface setup required!*

```
root:~# smc_rnics -a
FID  Power  PCI_ID        PCHID  Type           Port  PNET_ID
-------------------------------------------------------------
 80  1      0000:00:00.0  07c0   ISM            n/a   n/a
 81  0

root:~# smc_rnics -e 81
root:~# smc_rnics
FID  Power  PCI_ID        PCHID  Type           Port  PNET_ID
-------------------------------------------------------------
 80  1      0000:00:00.0  07c0   ISM            n/a   n/a
 81  1      0001:00:00.0  07c0   ISM            n/a   NET2
```

Works with any configuration!

Would require interface with identical PNET ID!

SMC-Dv2 – How to Use: Setup Verification

- **`smcd info`**[1]: Verify hardware and software support

```
root:~# smcd info
   Kernel Capabilities
   SMC Version:      2.0
   SMC Hostname:     tux
   SMC-D Features:   v1 v2
   SMC-R Features:   v1

   Hardware Capabilities
   SEID:             IBM-SYSZ-ISMSEID0000...
   ISM:              v1 v2
   RoCE:             v1
```

- **`smc_chk`**[1]: Live-test connectivity (will also report local setup issues), shipped with RHEL 8.5

```
root@t83lp76:~# smc_chk -S &
Server started on port 37373

root@t83lp76:~# smc_chk -C 127.0.0.1 -p 37373
Test with target IP 127.0.0.1 and port 37373
  Live test (SMC-D and SMC-R)
    Success, using SMC-D

root@t83lp76:~# smc_chk -C 192.168.5.47 -p 23
  Live test (SMC-D and SMC-R)
    Failed  (TCP fallback), reasons:
      Client:  0x03010000  Peer does not support SMC
```

Live test with local server, also provided by `smc_chk`

Live test with ssh server on remote

[1] *smc-tools* v1.5 or later

SMC-Dv2 – How to Use

1) Use pre-load library `libsmc-preload.so`

- Provided by *smc-tools*
- Intercepts existing applications' `socket()` calls
- Two ways to enable:
  a) Use `smc_run` (recommended)

```
$ smc_run <my_application>
```

  b) Enable through environment variable:

```
$ export LD_PRELOAD=libsmc-\
                    preload.so
```

- **Note**: Will not work with statically linked applications! (rare case)

2) Alternative: Re-compile the application

- SMC implemented as separate address family `AF_SMC`.
- In applications' `socket()` calls, replace `AF_INET` with `AF_SMC`, i.e.:

```
int s, ipv6 = 0;

s = socket(AF_SMC, SOCK_STREAM, ipv6);
```

- Unlikely to happen with users' applications

SMC-Dv2: How To Verify

- **What it does**:Provides summary of SMC enabled connections (successful and fallback)
- **Why you should care**: Can serve as basis for further optimizations to improve performance
- More details available, see option `--details`
- Supports data export in JSON format for further processing

```
root:~# smcd stats
SMC-D Connections Summary
  Total connections handled        152730
  SMC connections                  152730
  Handshake errors                      0
  Avg requests per SMC conn          813.1
  TCP fallback                          0

RX Stats
  Data transmitted (Bytes)   270311225427 (270.3G)
  Total requests                 61619256
  Buffer full                      114746 (0.19%)
            8KB    16KB    32KB    64KB   128KB   256KB   512KB  >512KB
  Bufs        0       2     140  2.103K       0       2       0  95.97K
  Reqs   54.07M       3  7.552M       0       0       0       0       0

TX Stats
  Data transmitted (Bytes)   271274963896 (271.3G)
  Total requests                 62565728
  Buffer full                           0 (0.00%)
  Buffer full(remote)               90038 (0.14%)
  Buffer too small                      0 (0.00%)
  Buffer too small(remote)              0 (0.00%)
            8KB    16KB    32KB    64KB   128KB   256KB   512KB  >512KB
  Bufs        0  2.384K     142       0       0       2       0       0
  Reqs   54.92M       3  7.552M       0       0       0       0       0

Extras
  Special socket calls                  0
```

Miscellaneous

Need something else for Linux & Virtualization?

Linux and KVM

(A) Use the Request for Enhancements (RFE) database:
- enter in your IBM ID
- select Brand "*Servers and System Software*"
- select Product "*Linux on System z*" (includes KVM)

(B) Reach to us at conferences, e.g.
SHARE GSE Conferences Tech. Univ VM Workshop

How the Linux Distro Partners handle requirements

Red Hat defined RFE process for customers

SUSE requirements can be submitted to their sales reps as well as using the "feedback" button at the bottom of the SUSE Linux Enterprise Server for IBM Z and LinuxONE web site

Canonical is handling requirements for Ubuntu through Launchpad: Open a bug, put requirement in title and tag with s390x

More information about Linux & KVM

- Official web site
  https://www.ibm.com/it-infrastructure/z/os/linux
- Linux & KVM (see Backup)
  Key Documentation Links
- Secure Execution & Compression (see Backup)
  Videos & books
- Enterprise Key Management for Linux (see Backup)
  Videos & books

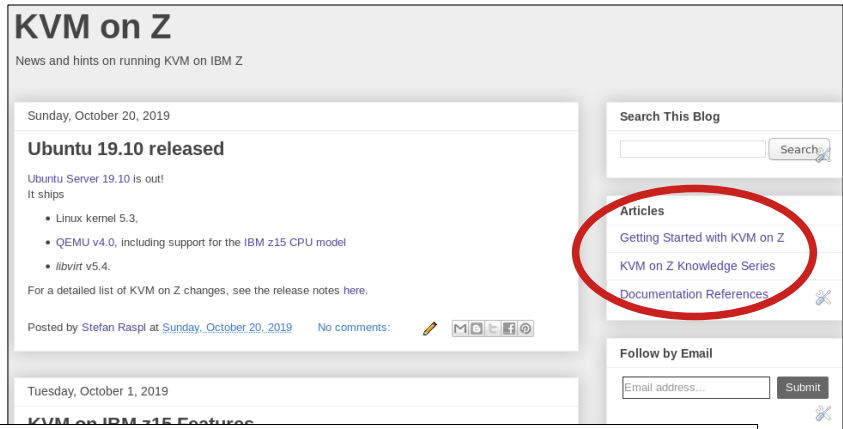
User forums

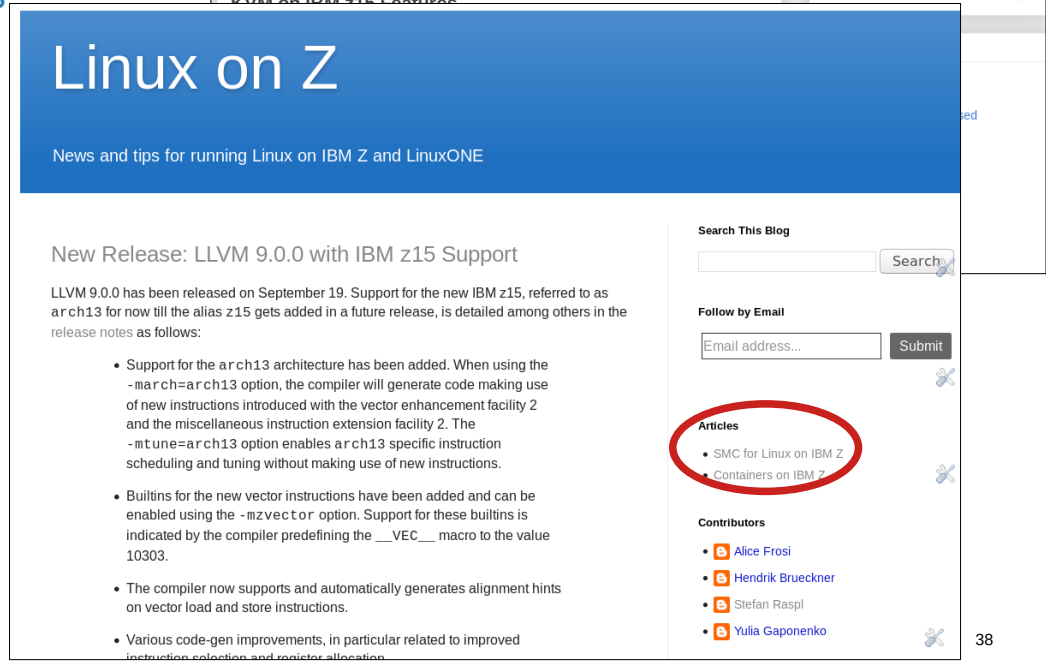
- Mailing lists at Maris college
- Linux on s390x forum at Open Mainframe Project

Staying Up-To-Date

Blogs

- Very latest news from the development team
  - KVM on Z: http://kvmonz.blogspot.com/
  - Linux on Z & containers: http://linux-on-z.blogspot.com/
- Focus primarily on upstream submissions, which will end up in Linux distributions later
- Also features in-depth articles on specific topics
- Provided by Linux & KVM on Z development teams





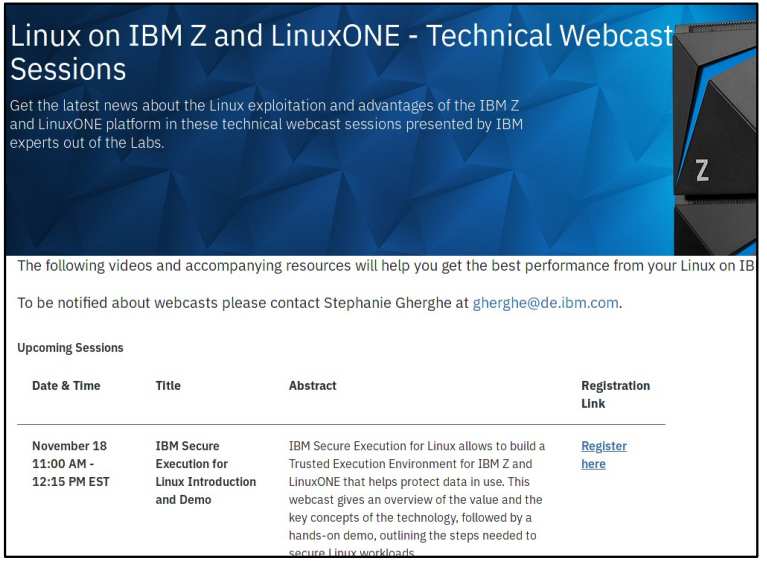
References

Documentation

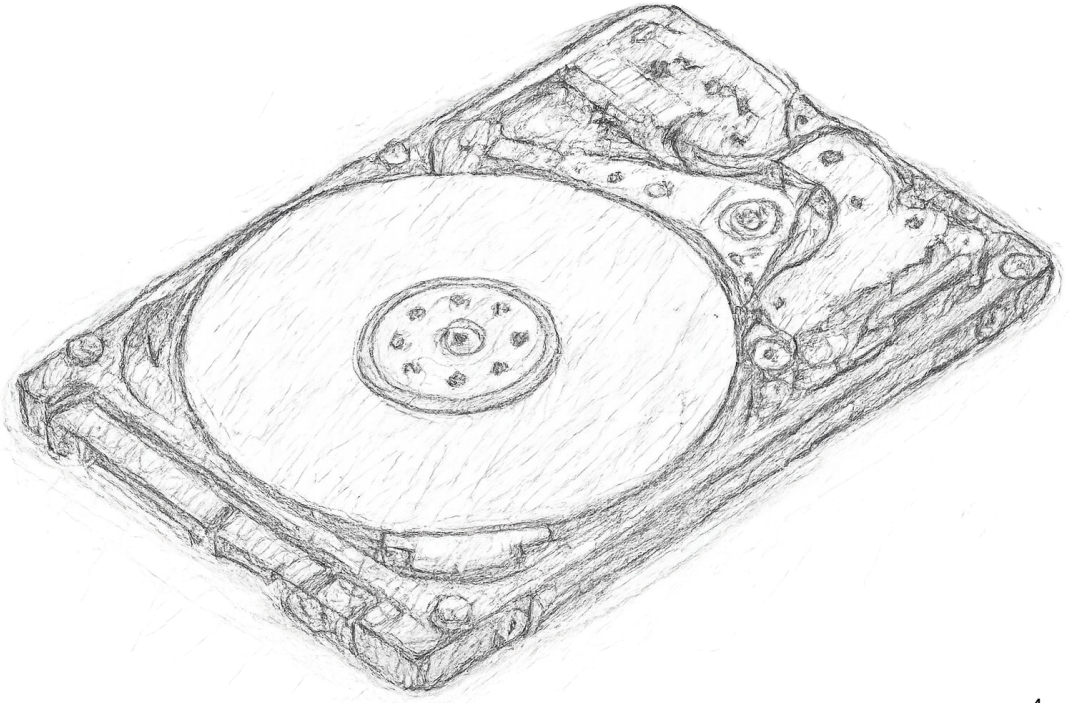
- Linux on Z and LinuxONE on IBM Documentation
  https://www.ibm.com/docs/en/linux-on-systems?topic=linux-z-linuxone
- Videos explainers
  https://www.ibm.com/docs/en/linux-on-systems?topic=linuxone-video-explainers
- Solution assurance
  https://www.ibm.com/docs/en/linux-on-systems?topic=linuxone-solution-assurance
- z/VM Education Roadmap
  https://www.vm.ibm.com/education/

Webcasts

- In-depth sessions right from the Linux on Z development team
- Recordings available
  https://ibm.biz/Linux-on-IBMzSystems-LinuxONE-Webcasts



Storage



Protected Key Crypto

- Data at rest
  - In Linux, the most popular method for end-to-end data at rest encryption is full volume encryption using the dm-crypt kernel component. dm-crypt reads encrypted sectors from a block device (disk, partition, or logical volume), decrypts the data in the sector, and writes it to the reading component (for example, into the page cache).
  - The challenge is to manage the cryptographic keys needed to open an encrypted
    - How can these keys be stored securely?
    - How can these keys be protected from being discovered while they are in
    - How can these keys be protected from being stolen by an intruder?
    - How can these keys be protected from discovery by a service engineer that

Protected Key Crypto...?
Hint only!!
Note: Book exists, see
https://public.dhe.ibm.com/software/dw/linux390/docu/l5n1dc02.pdf

  - To manage the challenge of storing cryptographic keys and associating these keys with the volume they encrypt, the Linux cryptsetup utility applies the Linux Unified Key Setup (LUKS) volume format. This format provides protection by passphrases and stores passphrase-protected volume keys in the header of the volume.
  - Protected keys support high performance AES cryptography using the acceleration of cryptographic operations provided by the IBM Z Central Processor Assist for Cryptographic Functions (CPACF)
  - Protected keys are volatile keys encrypted by the IBM firmware master key of an LPAR or a virtual machine. They are created by a specific instance of an LPAR or a z/VM® guest and can only be used by the instance of the operating system that created the protected key. Also, they are only valid as long as the LPAR or the virtual machine that generated the key is running. The operating system has no access to the plain values of protected keys, and protected keys are useless on any other system.
  - Therefore neither secure nor protected keys are of any use outside their system. Even if stolen, they cannot be used to decipher data stored in the storage system or in transit through the SAN from a system owned by adversary persons.
  - The PAES cipher is used by the infrastructure for protected volume encryption to enable the utilization of a protected key for the AES algorithm. The PAES cipher converts a secure key into a protected key for enhanced performance using the pkey module.

CPACF: How to Verify

- **How to use:**
- **What it does**:
  - Processor Activity Instrumentation (PAI) counters available with IBM z16 and LinuxONE 4 allow counting CPACF subfunctions
- **Why you should care**:
  - Check cipher usage by applications
  - Investigate use of weak crypto algorithms
  - Detailed information on ciphers and key lengths in use might be required for compliance reasons
- Linux provides counters for
  - Kernel space usage
  - Userspace usage
- See tool `cpacfstats` as shipped with `s390-tools 2.23` or later for further details

Usage

See tool `cpacfstats` from s390tools 2.23 or later

- New counter classes:
  - `pai_kernel` PAI kernel counters
  - `pai_user` PAI user space counters
  - to be used with enable (-e), disable (-d), reset (-r), and print (-p) options

Linux (on Z and LinuxONE) encryption of data-at-rest

e2e encryption

- dm-crypt: block device / full volume encryption
  - uses kernel crypto
  - granularity: disk partition / logical volume
- ext4 with encryption option: file system encryption
  - uses kernel crypto
  - granularity: file, directory, symbolic link
- Spectrum Scale (GPFS) with encryption option: file encryption
  - uses GSKit or Clic crypto libraries
  - granularity: file
- DB2 native encryption: data base encryption
  - uses GSKit crypto library
- NFS v4 with encryption option: encryption of file transport
  - uses kernel crypto

network encryption

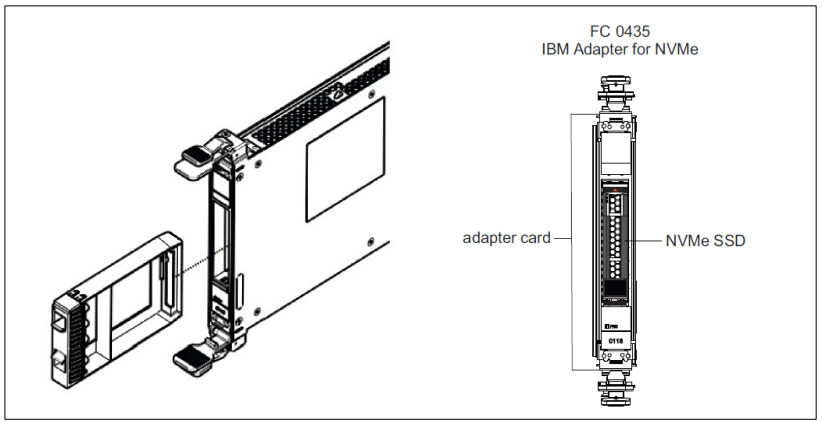
- SMB v3.1: encryption of file transport
  - uses kernel crypto

kernel crypto automatically uses CPACF for AES if the module aes_s390 is loaded

GSKit and latest versions of Clic use CPACF for AES

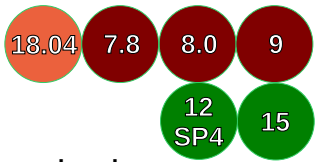
IBM LinuxONE support for NVMe drives

- **What it does:** Provide adapter for NVMe
  - Carrier card for industry standard U.2 NVMe drives
    - Common capacities up to 16 TB per drive
    - 1 drive per carrier, up to 16 cards per CEC
  - Available for IBM LinuxONE starting with Emperor II and Rockhopper II



- **Why you should care**:
  - Low-cost
  - Low-latency
  - High-throughput

- **NVMe drive characteristics**
  - PCI direct-attached (no SAN)
    - No cabling, switches, etc. required
  - No classic virtualization or shared access: can use one drive only in one LPAR/VM
  - KVM can split a single NVMes into multiple partitions/LVMs for multiple guests



- **Linux on Z support for NVMe**
  - Uses standard Linux NVMe driver
  - Always apply latest service levels!

Working with NVMe drives in Linux

Listing available devices

- Use `lspci` to show PCI device information
- Use `lsblk` to list NVMe block devices:

```
$ lsblk
NAME         MAJ:MIN RM  SIZE RO TYPE MOUNTPOINT
nvme0n1      259:0    0  3.6T  0 disk
├─nvme0n1p1  259:1    0  300M  0 part
├─nvme0n1p2  259:2    0   40G  0 part
├─nvme0n1p3  259:3    0  3.6T  0 part
└─nvme0n1p4  259:4    0    2G  0 part
```

Storing data

- Use NVMe for swap, root, boot* and data file systems (*=requires IPL support)
- Use of software RAID recommended

Management tools

- `nvme-cli`: query and manage NVMe device functions
- `zpcictl`: recovery and service actions

IPL support

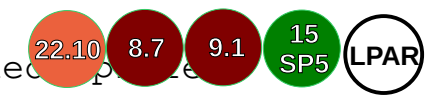
- Available on LinuxONE III with latest firmware
  - LPAR only
- KVM supports IPL via virtio-blk

- NVMe IPL support added with
  - Linux kernel v5.8
  - s390-tools v2.14
- Distribution installer support available

Load normal support

- Reduces boot time for large LPARs

Dump support



- Full tools integration, e.g. `makedu`
- Requires IBM z15 or later

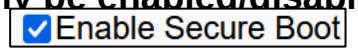
Secure Boot for ECKD DASD

LPAR z/VM

- **What it does:** Linux can boot from ECKD DASD in Secure Boot mode
- **Why you should care:** Secure Boot is a prerequisite for the NIAP certification, and deployment of Linux in environments with extra high security requirements
- **What you need:**
  - IBM z16 with GA1.5 firmware
  - **For Basic boot support:**
    - s390-tools v2.25

  - **For Reboot and dump support:**
    - s390-tools v2.26
    - Linux kernel v6.2

- **How to use it:** With the new support, Linux DASDs contain 2 types of boot loader:
  - CCW IPL: Standard boot
  - LD-IPL (“List-Directed IPL”): Supports Secure Boot
- **Note: Secure Boot can only be enabled/disabled on the HMC Load panel**



- **zipl will always install both boot loader types:**

```
$ zipl
Using config file '/etc/zipl.conf'
...
Preparing boot device for CCW- and LD-IPL: dasda (1234).
Done.
```

- **For reboot, IPL-type must be chosen manually**
  - `chreipl eckd` for DASD LD-IPL with Secure Boot support
  - `chreipl ccw` for DASD CCW-IPL with standard boot

```
$ chreipl eckd 0.1.1002
Re-IPL type: eckd
Device:      0.1.1002
bootprog:    0
br_chr:      auto
Bootparm:    ""
Loadparm:    ""
clear:       0
```

```
$ chreipl ccw 0.1.1002
Re-IPL type: ccw
Device:      0.1.1002
Loadparm:    ""
clear:       0
```

Multi-Path Re-IPL

- **What it does**: Keeps re-IPL path up-to-date with working path for next re-IPL
- **Why you should care**: Protects *vastly* against re-IPL issues due to IPL path becoming unavailable.
- **The `chreipl-fcp-mpath` toolset monitors udev events about paths to the re-IPL volume**
- **If currently configured FCP re-IPL path becomes unavailable, re-configures the FCP re-IPL settings alternative operational path to same volume**
- **Thus, re-IPL from an FCP-attached SCSI volume can be successful despite path failures on a running Linux instance if at least one path to the re-IPL volume remains operational**
- **See man-page for `chreipl-fcp-mpath` for further details**

```
# Activate by installing respective package
$ apt install s390-tools-chreipl-fcp-mpath

$ lsreipl
Re-IPL type: fcp
WWPN:        0x5005076309005430
LUN:         0x4018401600000000
Device:      0.0.1700
bootprog:    0
br_lba:      0
Loadparm:    ""
Bootparms:   ""
# Cable Pull/Switch Port Toggle/Path Goes Away

$ journalctl --boot --identifier=chreipl-fcp-mpath
May 02 10:04:42 t3545003 chreipl-fcp-mpath[46089]:\
Changed re-IPL path to: 0.0.1740:0x5005076309045430\
:0x4018401600000000.

$ lsreipl
Re-IPL type: fcp
WWPN:        0x5005076309045430
LUN:         0x4018401600000000
Device:      0.0.1740
bootprog:    0
br_lba:      0
Loadparm:    ""
Bootparms:   ""
```

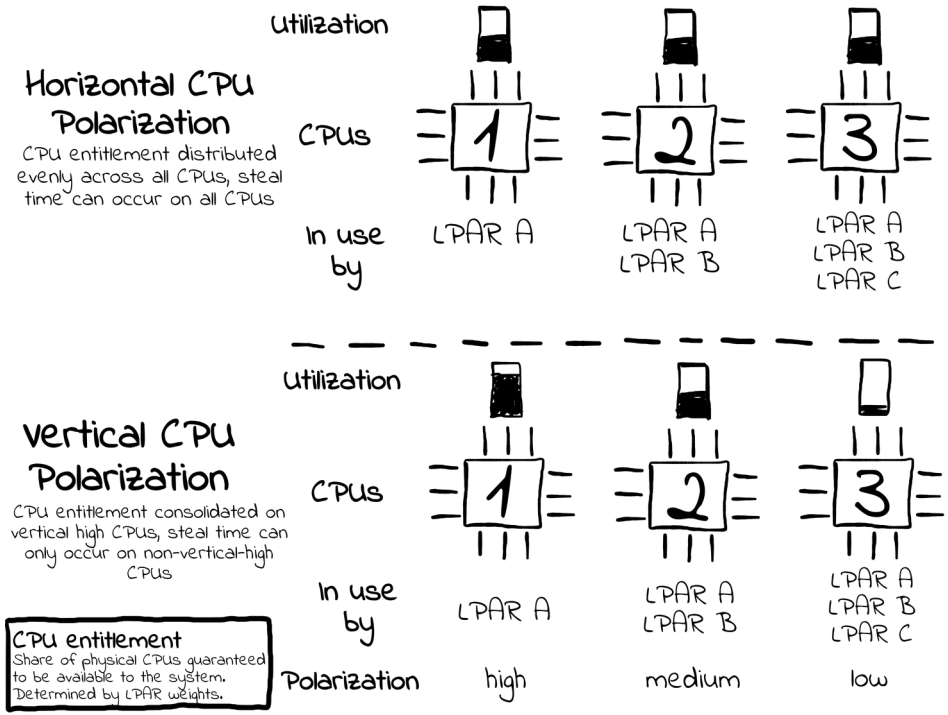
Fig.1: *Sample output*

Linux / Misc

HiperDispatch Support aka Vertical CPU Polarization

- **What it does:** Prioritize process scheduling to CPUs with more consistent processing guarantees to avoid steal time
- **Why you should care:** Can yield substantial performance improvements for CPU-intensive workloads on highly utilized CECs
- Platform differentiates between *vertical high*, *medium* and *low* IFLs, with varying capacity grants
- Basically no steal time on *vertical high* IFLs
- Modifies the scheduler to prefer *vertical highs* and *mediums* for CPU-intensive workloads
- Workloads running large numbers of small tasks might perform better with horizontal CPU polarization
- **How to use:**
  - Enabled by default
  - Use sysctl `s390.hiperdispatch` to enable or disable:

```
sysctl -w s390.hiperdispatch=[0|1]
```



Tunables:

- `/sys/devices/system/cpu/hd_steal_threshold`
- `/sys/devices/system/cpu/hd_delay_factor`
  Steal time evaluation period. Reducing this value improves responsiveness to changes in workload behavior. Increasing it delays reaction to sudden changes in steal time.

Tag Legend

- Supported distributions
  - x.y for SUSE SLES <X> Service Pack <Y>, e.g. 15 SP6 for SLES15 SP6
  - x.y for RHEL <x> Update <y>, e.g. 9.4 for RHEL9.4
  - x.y for Ubuntu x.y, e.g. 24.04 for Ubuntu 16.04 LTS
- Suppored environments
  - LPAR usable for Linux virtual servers running in LPAR
  - z/VM usable for guests running on z/VM
  - KVM usable for guests running on KVM

Disclaimer for IBM Spyre™ Accelerator

The IBM Spyre™ AI Accelerator will not be available with IBM general availability. The IBM Spyre AI Accelerator is currently expected to be available in 4Q of 2025. Any capabilities discussed in this presentation with respect to the IBM Spyre AI Accelerator will not be enabled by until these accelerator cards are installed in the system.

Citations/Claims/Disclaimers

1.DISCLAIMER: IBM internal data based on measurements and projections was used in calculating the expected value. Necessary components include IBM LinuxONE Emperor 5; IBM z/VM V7.3 systems or above collected in a Single System Image, each running RHOCP 4.14 or above; IBM Operations Manager; GDPS 4.6 or above for management of data recovery and virtual machine recovery across metro distance systems and storage, including Metro Multi-site workload and GDPS Global; and IBM DS8000 series storage with IBM HyperSwap. A MongoDB v4.4 workload was used. Necessary resiliency technology must be enabled, including z/VM Single System Image clustering, GDPS xDR Proxy for z/VM, and Red Hat OpenShift Data Foundation (ODF) 4.14 or above for management of local storage devices. Application-induced outages are not included in the above measurements. Other configurations (hardware or software) may provide different availability characteristics.

2. The Cost of a Data Breach Report 2024, by Ponemon Institute and analyzed by IBM https://www.ibm.com/downloads/documents/us-en/107a02e94948f4ec

3. X-Force Threat Intelligence Index 2024, https://www.ibm.com/downloads/documents/us-en/10c31775c0d40a37

4. Dimple Ahluwalia, Gerry Parham, Kevin Skapinetz, (August 2024), Securing generative AI: What matters now. IBM Institute for Business Value: Full Data and insights deck. https://w3.ibm.com/services/lighthouse/documents/218436?ref=bmwiz

5. DISCLAIMER: Performance result is extrapolated from IBM® internal tests running on IBM Systems Hardware of machine type 9175. The Acme Air microservice benchmark (https://github.com/blueperf/acmeair-mainservice-java) was deployed on Red Hat® OpenShift® Container Platform (RHOCP) 4.17. The 3 RHOCP Compute nodes ran 3 Acme Air instances in parallel, each driven remotely from Apache JMeter™ 5.2.1 with 128 parallel users. IBM Systems Hardware configuration: 8 LPARs in total with 21 dedicated and 4 shared cores (SMT). 3 LPARs running RHOCP Compute nodes each with 7 dedicated cores (SMT), 64 GB memory and DASD storage. 5 LPARs each with 4 shared cores (SMT), 16 GB memory and DASD storage, providing 3 RHOCP Management nodes and 2 RHOCP Infrastructure nodes. The Network adapters were dedicated for NETH on Linux. Results may vary.

6. DISCLAIMER: Performance result is extrapolated based on IBM® internal tests running on IBM Systems Hardware of machine type 9175. The flexible I/O tester benchmark (fio 3.35) ran with 128 parallel threads using read-only and read-write (70:30) operations on 8x 40 GB files each on a separate FCP attached LUN. IBM Systems Hardware configuration: 1 LPAR running Red Hat® Enterprise Linux® 9.4 (upstream Kernel level 6.13 with aes-256-xts exploitation patch - Commit ID 80625b670312e74512d65b19e9470184386ab265) with 13 dedicated cores (SMT), 64 GB memory, 1 FCP Express32G adapter with 4 ports connected. 1 IBM FlashSystem® 9500 (FS9500) with 8 ports connected, providing 8 LUNs of 320 GB total size, equally distributed across the two node canisters. Each luks2 encrypted LUN is connected through 8 paths and formatted with an ext4 file system. Cores utilization targeted below 50%. Results may vary.

7. Michael Osborne, Katia Moskvitch, Jennifer Janechek (August 2024) , NIST's post-quantum cryptography standards are here. https://research.ibm.com/blog/nist-pqc-standards

8. 6 blind spots tech leaders must reveal. IBM Institute for Business Value. https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/cxo

9. Eren Çam, Zoe Hungerford, Niklas Schoch, Francys Pinto Miranda, Carlos David Yáñez de León. (2024), Electricity 2024: Analysis and forecast to 2026. https://iea.blob.core.windows.net/assets/6b2fd954-2017-408e-bf08-952fdd62118a/Electricity2024-Analysisandforecastto2026.pdf

10. Sreejit Roy, Nalini Manuru, Charbak Roy, Lisa Fisher, Jacob Dencik, (January 2025) . IT Sustainability at a crossroads: Choosing a future of responsible computing. https://w3.ibm.com/services/lighthouse/spotlight/documents/225251?listId=ytmdo

IBM Spyre™ Accelerator is
in tech preview until Q42025

Citations/Claims/Disclaimers

11.CLAIM: IBM internal tests simulating a complete IT solution running containerized WebSphere Liberty and EDB Postgres workloads, show that a single IBM LinuxONE Emperor 5 Max 136 can do the work of up to 2,944 cores of the compared x86 solution.
DISCLAIMER: IBM® internal performance tests for the core consolidation study targeted a comparison of the following servers. IBM Machine Type 9175 MAX 136 system consisting of three CPC drawers containing 136 configurable processor units and six I/O drawers to support both network and external storage. The x86 solution used a commercially available enterprise server with two 5th generation Intel® Xeon® Platinum 8592+ processors, 64 cores per CPU. Both solutions had access to the same storage. The workloads consisted of a containerized online transaction processing (OLTP) WebSphere Liberty v25 application running on Red Hat OpenShift Container Platform (OCP) v4.17, and an EDB Postgres for Kubernetes v1.25 on the same OCP cluster simulating core online banking functions. Both solutions used Red Hat Enterprise Linux v9.5 and KVM. Results may vary.

The test results were extrapolated to a typical, complete customer IT solution that includes isolated from each other production and non-production IT environments. TCO included software, hardware, energy, network, data center space, and labor costs. On the IBM z17 side the complete solution requires one IBM z17 Type 9175 MAX 136, and on x86 side, the complete IT solution requires 23 compared servers.

12. DISCLAIMER: IBM internal performance tests for the core consolidation study compared an IBM Machine Type 9175 Max136 with 136 configurable processor units with an x86 solution that used a commercially available enterprise server with two 5th gen Intel Xeon Platinum 8592+ processors and 64 cores per CPU. Workloads consisted of a containerized OLTP WebSphere Liberty v25 application running on Red Hat OCP v4.17 and an EDB Postgres for Kubernetes v1.25 on the same OCP cluster. Both solutions used Red Hat Enterprise Linux v9.5 and KVM. Test results were extrapolated to a typical, complete customer IT solution that included production and non-production IT environments isolated from each other. The IBM Machine Type 9175 solution required one Max136 and the x86 solution required 23 compared servers. Results may vary.

13.DISCLAIMER: IBM® internal performance tests for the core consolidation study targeted a comparison of the following servers. IBM Machine Type 9175 MAX 136 system consisting of three CPC drawers containing 136 configurable processor units and six I/O drawers to support both network and external storage. The x86 solution used a commercially available enterprise server with two 5th generation Intel® Xeon® Platinum 8592+ processors, 64 cores per CPU. Both solutions had access to the same storage. The workloads consisted of a containerized online transaction processing (OLTP) WebSphere Liberty v25 application running on Red Hat OpenShift Container Platform (OCP) v4.17, and an EDB Postgres for Kubernetes v1.25 on the same OCP cluster simulating core online banking functions. Both solutions used Red Hat Enterprise Linux v9.5 and KVM. Results may vary.

The test results were extrapolated to a typical, complete customer IT solution that includes isolated from each other production and non-production IT environments. TCO included software, hardware, energy, network, data center space, and labor costs. On the IBM z17 side the complete solution requires one IBM z17 Type 9175 MAX 136, and on x86 side, the complete IT solution requires 23 compared servers.

14. Yunke Wang, Yanxi Li, Chang Xu AI Scaling: From Up to Down and Out, Feb 2025; https://arxiv.org/html/2502.01677v1

15.Bieth Stackpole. AI has high data center energy cost but there are solutions. (January 7, 2025) https://mitsloan.mit.edu/ideas-made-to-matter/ai-has-high-data-center-energy-costs-there-are-solutions

16. Cost of a Data Breach 2024, IBM and Ponemon Institute 2024, https://www.ibm.com/downloads/documents/us-en/107a02e94948f4ec

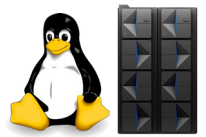
Citations/Claims/Disclaimers

17.CLAIM: The IBM Integrated Accelerator for AI on IBM LinuxONE Emperor 5 at full utilization is designed to process up to 24 trillion operations per second (TOPS) shared across all cores on the chip.

DISCLAIMER: Result is the maximum theoretical number of trillion operations per second (TOPS) in 8bit precision that can be executed by a single IBM Integrated Accelerator for AI. Cores are running at 5.5GHz and have one IBM Integrated Accelerator for AI per chip. The IBM Integrated Accelerator for AI consists of 2 corelets, each with an array of 64 tensor cores capable of executing 4 integer-multiply-add operations (IMA) 8-way SIMD with no sparsity.

IBM Spyre™ Accelerator is
in tech preview until Q42025

Pause-Less Garbage Collection



LPAR z/VM KVM

- **Support for the *Guarded Storage Facility* (kernel 4.12)**
  - Designed to improve the performance of Java applications that are constrained by *Garbage Collection* (GC)
  - Up to 64 regions of memory can be marked as being *guarded*
  - Reading a pointer with the new `LGG` or `LLGFSG` instruction will do a range check on the loaded value and automatically invoke a user space handler if one of the guarded regions is affected
  - Prerequisites:
    - IBM z14 or LinuxONE II or later
    - Java 8 SR5