zCX Performance Considerations

Michael Fitzpatrick – IBM z/OS Communications Server, mfitz@us.ibm.com
Nicholas Matsakis – IBM Poughkeepsie, z/OS Development, Matsakis@us.ibm.com

November 2020
Session 1AE

Disclaimers

All performance information was determined on dedicated hardware in a controlled environment. Actual results may vary.

Performance information is provided “AS IS” and no warranties or guarantees are expressed or implied by IBM.

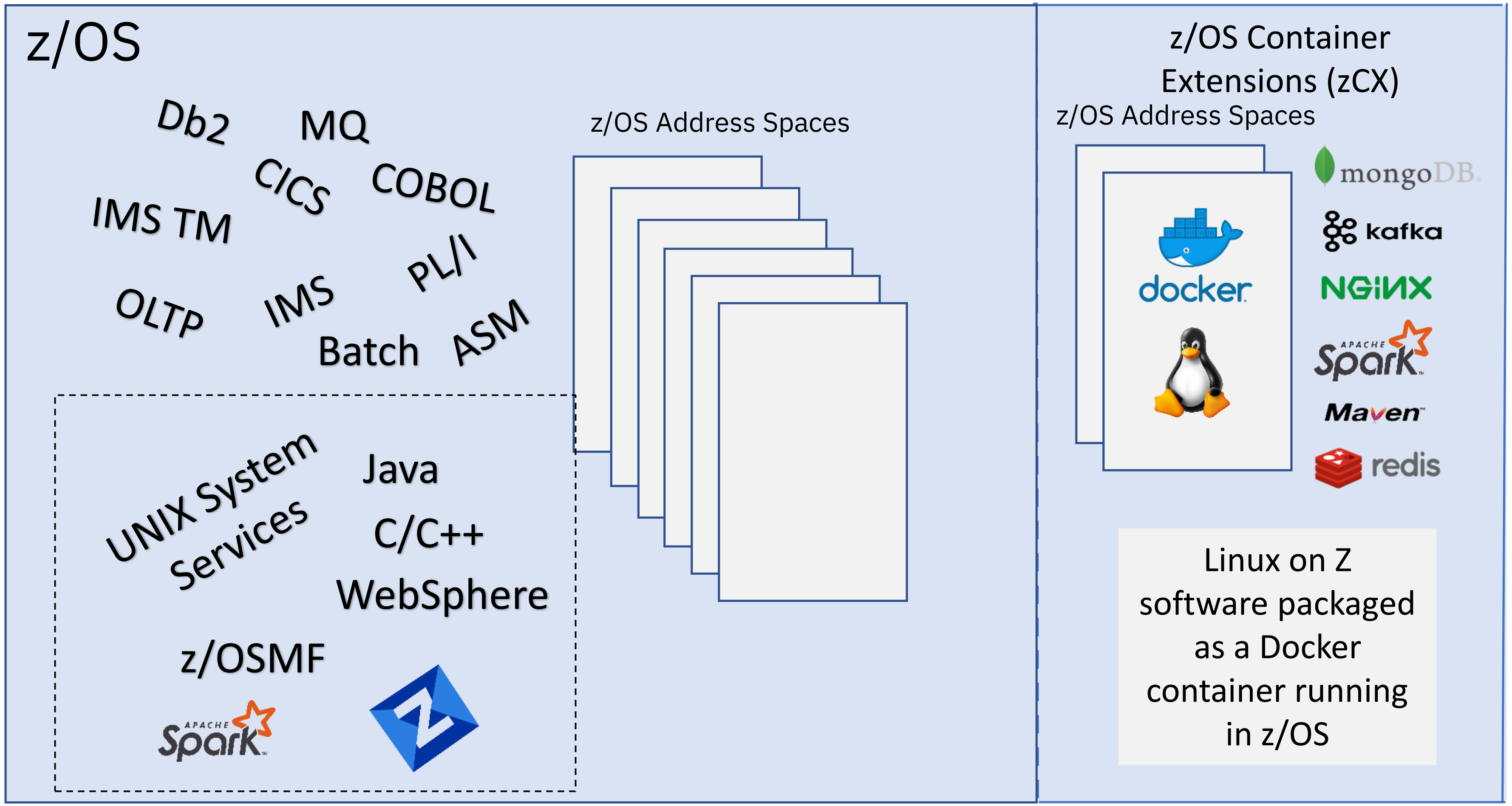
IBM statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion.

Agenda

- z/OS Container Extensions (zCX) Overview
- zCX Performance Considerations
  - Configuration: Capacity Planning, Page Frame Size, WLM, …
  - Network Configuration
- Docker Performance Monitoring
- Reference Material

z/OS Container Extensions (zCX) Overview

Expanding the z/OS Software Ecosystem

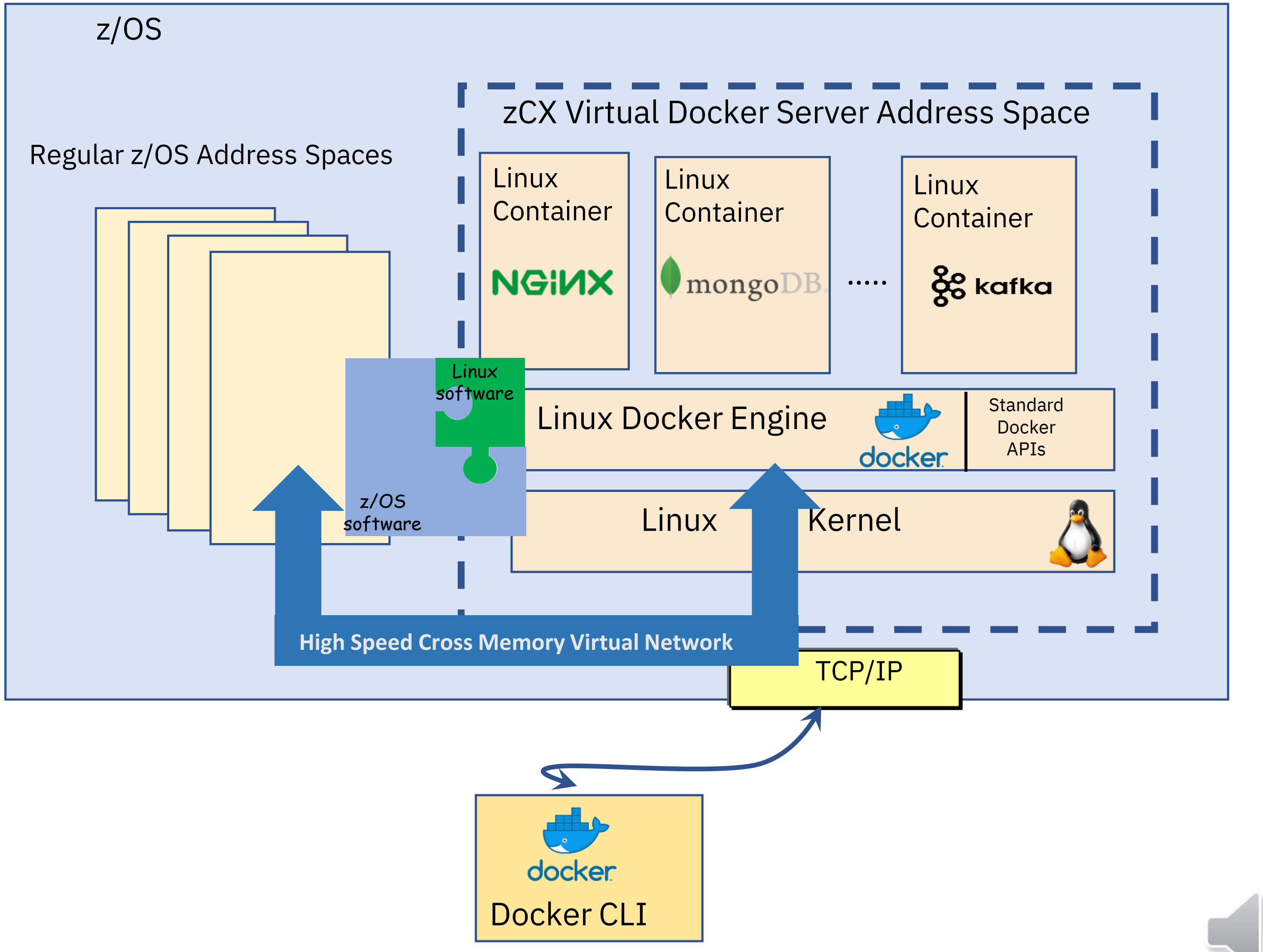


Linux on Z software packaged as a Docker container running in z/OS

- Traditional z/OS workloads, middleware, subsystems and programming languages
- UNIX System Services provided z/OS with a UNIX personality enabling porting of applications and new programming languages to the platform
- z/OS Container Extensions (zCX) provides the next big evolution – unmodified Linux on Z Docker images running inside z/OS

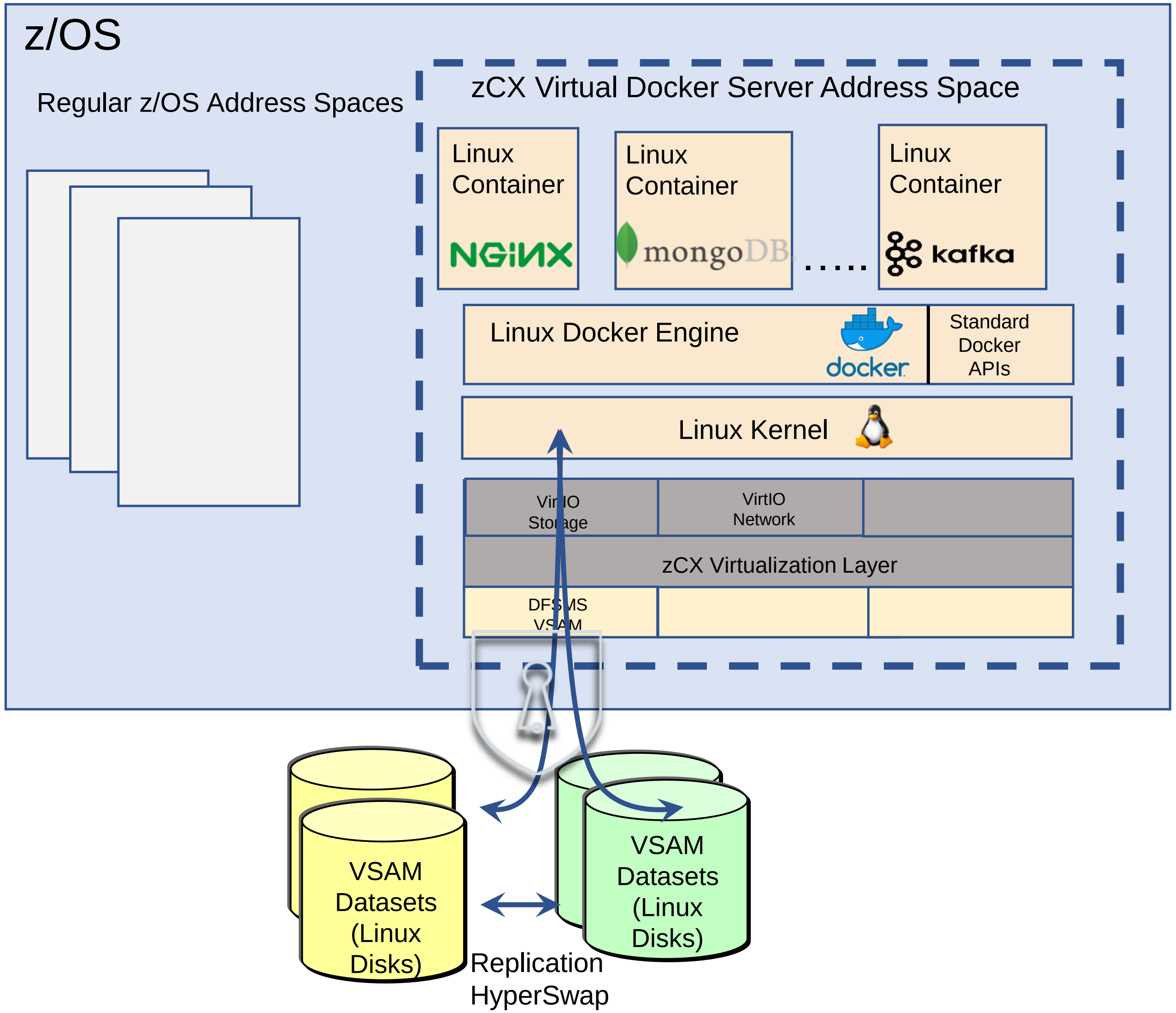
zCX – A turn-key Virtual Docker Server Software Appliance

- Pre-packaged Linux Docker appliance
  - Provided and maintained by IBM
  - Provisioned using z/OSMF workflows
- Provides standard Docker interfaces
  - Supports deployment of any software available as a Docker image for Linux on Z
  - Communications with native z/OS applications over high-speed virtual IP network
  - No z/OS skills required to develop and deploy Docker Containers
- No Linux system administration skills required
  - Interfaces limited to Docker CLI
  - No direct access to underlying Linux kernel
- Managed as a z/OS address space
  - Multiple instances can be deployed in a z/OS system
  - Managed using z/OS Operational Procedures
  - zCX workloads are zIIP eligible



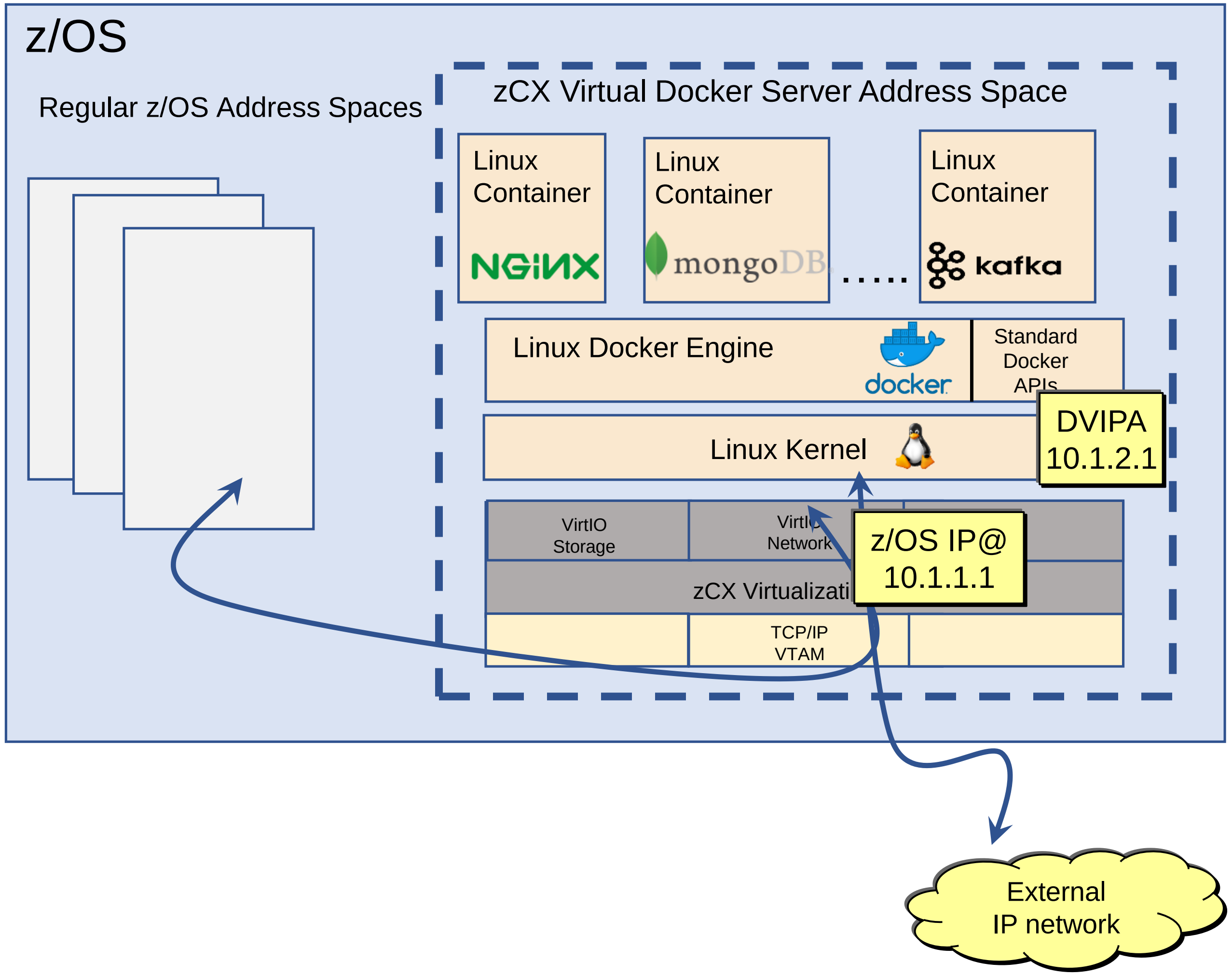
IBM zCX – z/OS Storage Integration

- z/OS Linux Virtualization Layer
  - Allows virtual access to z/OS Storage
  - Using virtio Linux interfaces
    - Stable, well defined interfaces used to virtualize Linux
  - Allows us to support unmodified, open source Linux on Z
- Linux storage/disk access
  - Uses z/OS owned and managed VSAM datasets
  - Leverages latest I/O enhancements
  - Built-in host-based encryption
  - Take advantage of existing replication technologies and HyperSwap



IBM zCX – z/OS Network Integration

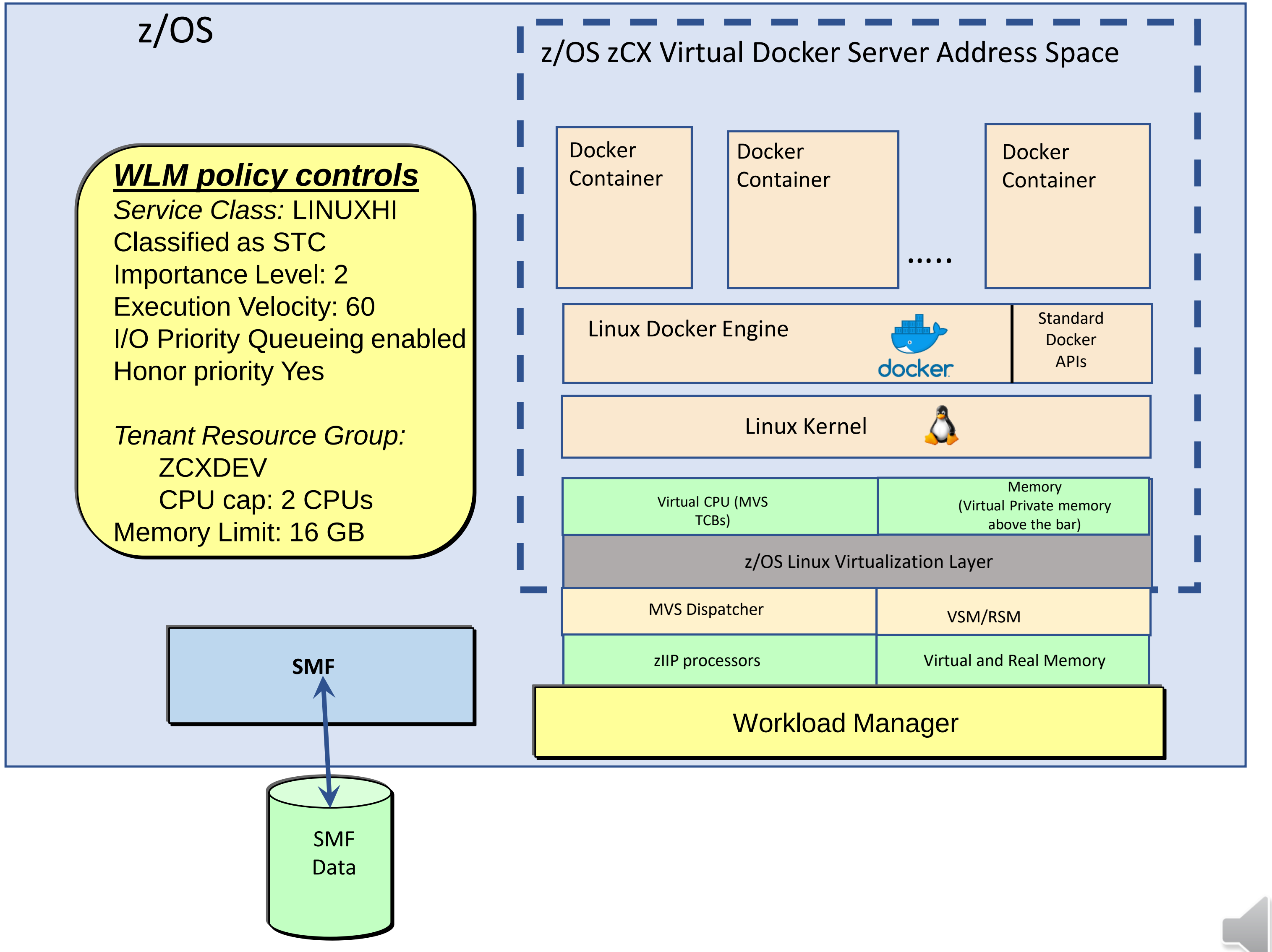
- z/OS Linux Virtualization Layer
  - Allows virtual access to z/OS Network
  - Using virtio Linux interfaces
    - Stable, well defined interfaces used to virtualize Linux
- Linux network access via high speed virtual *SAMEHOST* link to z/OS TCP/IP protocol stack
  - Each Linux Docker Server represented by a z/OS owned, managed and advertised Dynamic VIPA (DVIPA)
    - Allows restart of a CX instance in another system in the sysplex
  - Provide high performance network access across z/OS applications and Linux Docker containers – leveraging cross memory
    - All communications between zCX containers and z/OS applications over TCP/IP
  - External network access via z/OS TCP/IP
    - z/OS IP filters to restrict external access



zCX Performance Considerations: Configuration, Capacity Planning, WLM…

IBM zCX - CPU, Memory and Workload Management

- Memory Management
  - Provisioned per zCX Docker Server address space
  - Private, above the 2GB bar fixed memory
  - Managed by VSM, RSM
- CPU Management
  - Virtual CPUs provisioned to each zCX Docker Server address space
    - Each virtual CPU is a dispatchable thread (i.e. MVS TCB) within the address space
    - zIIP CPU access via MVS dispatcher
  - A zCX instance can host multiple Docker Container instances
- Normal WLM policy and resource controls extend to zCX Docker Server address spaces
  - Service Class association, goals, and importance levels
    - No transactional level controls
  - Tenant Resource Group association
    - Optional caps for CPU and real memory
- Normal SMF data available
  - SMF type 30, 72, etc.
  - Enables z/OS performance management and capacity planning



WLM policy controls

Service Class: LINUXHI
Classified as STC
Importance Level: 2
Execution Velocity: 60
I/O Priority Queueing enabled
Honor priority Yes

Tenant Resource Group:
ZCXDEV
CPU cap: 2 CPUs
Memory Limit: 16 GB

IBM zCX – Guest CPU Usage

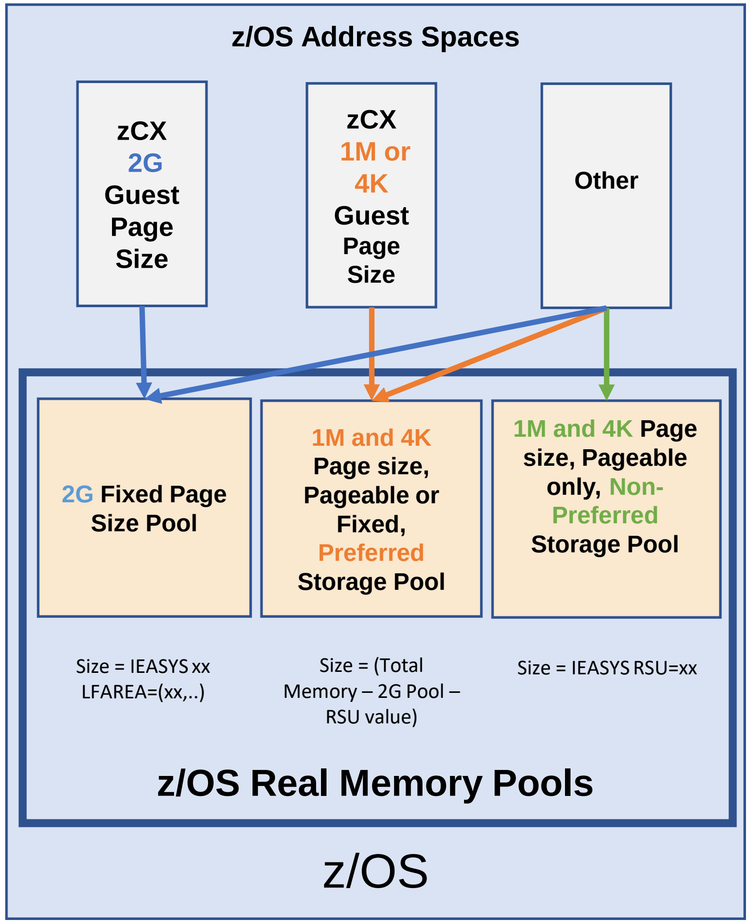
- Choosing the number of Appliance Virtual CPUs
  - The Virtual CPUs are what Linux uses
    - Docker provides controls to limit individual container CPU and Memory usage
  - Reflective of the amount of CPU power required during peak times to meet the required goals
  - Don’t define more than the number of zIIP threads (SMT-2 is fine) on the LPAR

- How much zIIP Eligibility?
  - Internal IBM workloads in a controlled environment have shown up to 98% zIIP eligibility
    - Rule of thumb is 95% of a workload will be zIIP eligible but results will vary
  - Plan for the increased CPU utilization to prevent impacting existing zIIP eligible workloads
    - See Martin Packer's "zIIP Capacity And Performance“ presentation for capacity planning guidance
  - IEAOPTxx IIPHONORPRIORITY(YES) allows zIIP eligible work to run on GCPs when zIIP capacity is not available:
    - GCP offload increases software cost
    - Review RMF APPL% IIPCP to determine zIIP eligible work that ran on GCPs
    - WLM Resource groups can be used to limit zIIP offload to GCPs but will cause delays

- What is zIIP Eligible?
  - All virtual CPUs
    - Some MVS processing must be done on GCPs
  - SRBs processing inbound TCP/IP traffic when IWQ (Inbound Work Queuing) is used

- What is not zIIP Eligible?
  - SRBs for I/O completion
    - Not CPU intensive anyway
  - SRBs processing inbound TCP/IP traffic when IWQ (Inbound Work Queuing) not used
    - Not recommended

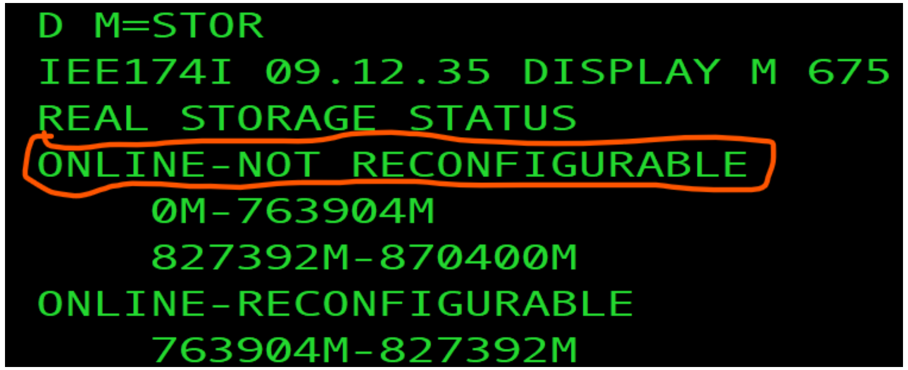
IBM zCX – Guest Memory Considerations


z/OS Address Spaces
zCX 2G Guest Page Size
zCX 1M or 4K Guest Page Size
Other
2G Fixed Page Size Pool
1M and 4K Page size, Pageable or Fixed, Preferred Storage Pool
1M and 4K Page size, Pageable only, Non-Preferred Storage Pool
Size = IEASYS xx LFAREA=(xx,..)
Size = (Total Memory – 2G Pool – RSU value)
Size = IEASYS RSU=xx
z/OS Real Memory Pools
z/OS

- Memory size is a function of what the containers require plus 1GB
  - See the zCX documentation for rules of thumb to calculate the real and swap memory sizes
  - Provide enough memory so Linux does not page
  - Docker has control limits for container CPU and Memory
- Fixed High Private (>Bar) is used for guest memory
  - >Bar virtual storage MEMLIMIT control does not apply
  - Prior to APAR OA59573, only a 4k real page size can be used
  - Only preferred (non-reconfigurable) real storage from the 2G fixed page size or 1M/4K preferred page pool can be used
- z/OS will start paging and swapping out address spaces when the fixed storage threshold for non-2G page storage is reached
  - Can negatively impact other work on the system
  - Plan for the preferred storage that will be required
    - The appliance will not start if it will cause the fixed storage threshold to be reached
    - WLM Resource Group Memory Pools can be used to limit address space fixed storage consumption
  - IEAOPTxx OPT parameters related to the fixed storage threshold
    - IRA405I(2) controls the fixed percentage of the non-2G preferred and reconfigurable storage when message IAR405I is issued
    - MCCFXTPR controls the fixed percentage of the non-2G storage when the system will start to take action to control fixed large consumers
      - The default is 80% fixed of all non-2G preferred and reconfigurable (non-preferred) storage

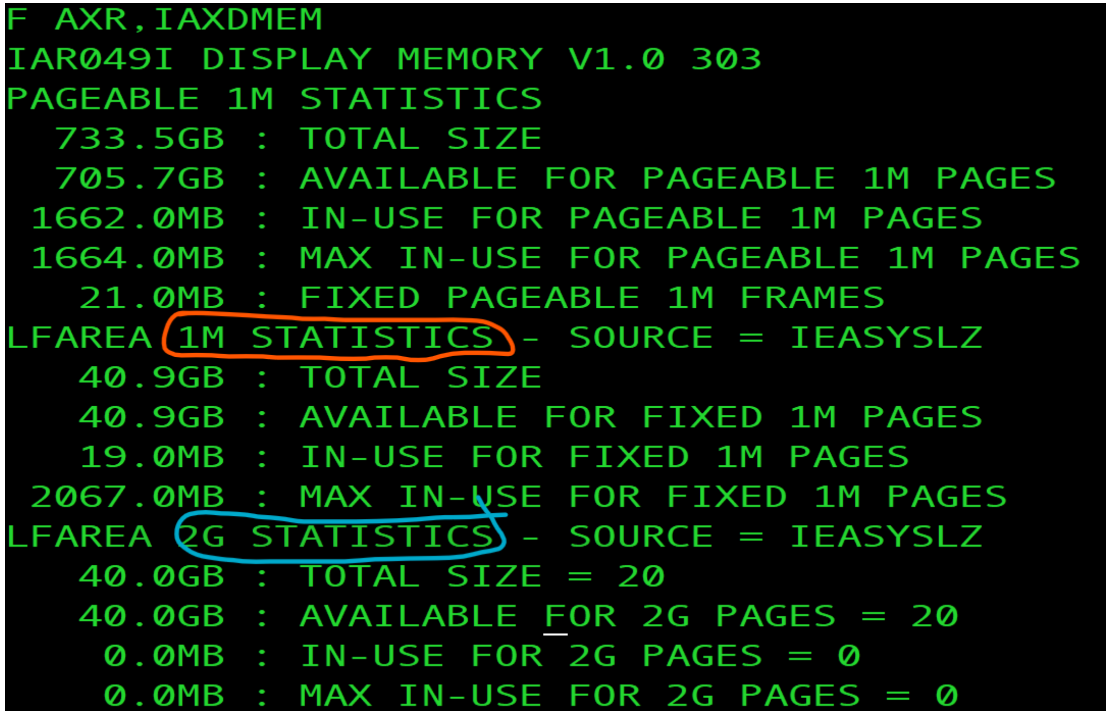
IBM zCX – Guest Memory Considerations

- The D M=STOR and F AXR,IAXDMEM commands can be used to display the RSU value, storage pool values, and the amount being used:



```
 D M=STOR
 IEE174I 09.12.35 DISPLAY M 675
 REAL STORAGE STATUS
 ONLINE-NOT RECONFIGURABLE
     0M-763904M
     827392M-870400M
 ONLINE-RECONFIGURABLE
     763904M-827392M
```

- 788GB of Non-Reconfigurable preferred storage



```
F AXR,IAXDMEM
IAR049I DISPLAY MEMORY V1.0 303
PAGEABLE 1M STATISTICS
  733.5GB : TOTAL SIZE
  705.7GB : AVAILABLE FOR PAGEABLE 1M PAGES
 1662.0MB : IN-USE FOR PAGEABLE 1M PAGES
 1664.0MB : MAX IN-USE FOR PAGEABLE 1M PAGES
   21.0MB : FIXED PAGEABLE 1M FRAMES
LFAREA 1M STATISTICS - SOURCE = IEASYSLZ
   40.9GB : TOTAL SIZE
   40.9GB : AVAILABLE FOR FIXED 1M PAGES
   19.0MB : IN-USE FOR FIXED 1M PAGES
 2067.0MB : MAX IN-USE FOR FIXED 1M PAGES
LFAREA 2G STATISTICS - SOURCE = IEASYSLZ
   40.0GB : TOTAL SIZE = 20
   40.0GB : AVAILABLE FOR 2G PAGES = 20
    0.0MB : IN-USE FOR 2G PAGES = 0
    0.0MB : MAX IN-USE FOR 2G PAGES = 0
```

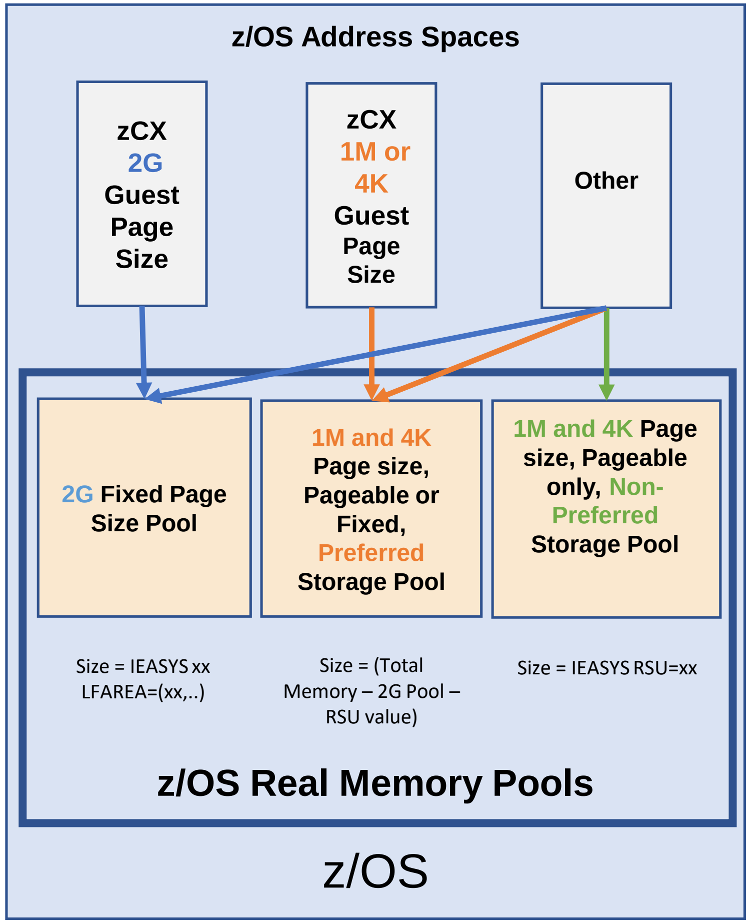
- 20GB of 2G fixed pages all of which are in use
- 40.9GB of 1M fixed pages of which 19MB are in use

IBM zCX – Guest Memory Page Frame Size (OA59573)

- zCX APAR **OA59573** available as of Sept 30, 2020
  - Provides the ability to back the Linux memory with 2G fixed, 1M fixed, or 4K fixed pages
    - Larger pages reduce Translation Lookaside Buffer misses and Translation Table sizes
    - 2G and 1M pages save about 8MB of Translation Table space for every 2GB of memory
    - Do not get Linux Hugepages mixed up with this support. Linux Hugepages are currently not supported.
    - Most workloads showed significant benefits. However, rarer small memory footprint workloads may show little benefit
  - Has additional performance improvements
  - Internal benchmarks performed in a dedicated controlled environment showed the following improvements compared to a zCX without OA59573 applied:

Page Size	**% ITR Range**	**% ETR Range**
2G	1- 13	>0 – 6
1M	1 - 10	>0 - 5
4k	>0 - 1	0 - 2

IBM zCX – Guest Memory Page Frame Size (OA59573)


z/OS Address Spaces
zCX 2G Guest Page Size
zCX 1M or 4K Guest Page Size
Other
2G Fixed Page Size Pool
1M and 4K Page size, Pageable or Fixed, Preferred Storage Pool
1M and 4K Page size, Pageable only, Non-Preferred Storage Pool
Size = IEASYS xx LFAREA=(xx,..)
Size = (Total Memory – 2G Pool – RSU value)
Size = IEASYS RSU=xx
z/OS Real Memory Pools
z/OS

How to choose a page size:

- 2G fixed pages
  - Best performance:
    - Reduces TLB misses and page table storage as one 2GB page contains 524,288 4k pages and 2048 1M pages
  - Least flexible
    - The storage is pre-allocated at IPL time via the 2G LFAREA parameter and cannot be used for any other storage pool
    - If storage consumption is an issue, then it may not be the best choice for instances that come and go as others may not be able to use the 2G memory
    - Cannot be used when z/OS is a z/VM guest
- 1M fixed frames
  - Improved performance over 4K but noticeably less than 2G
    - Reduces TLB misses and page table storage as a 1M page contains 256 4K pages
    - On z/VM, provides a dramatic performance improvement over 4k frames and is highly recommended
  - Good flexibility
    - Good choice for appliances that come and go as storage can be reused as 4k if needed
    - 1M LFAREA parameter is only a maximum value as there is no dedicated 1M page pool
- 4K frames
  - Worst performance - not recommended as a first choice unless a sandbox appliance
  - Best availability - add as a second choice in case your first choice is unavailable
- Recommendations
  - Use 2G pages for appliances that are always up or when reusing memory is not an issue
  - Pick a combination of sizes starting with your first choice and the system will use the first one that is available (i.e. 2G, 4K)
  - Automate for cases where the best size was not available but should be

IBM zCX – Latest Service

- zCX is delivering functionality both within and in-between z/OS releases

- Make sure you have the latest service before you start, as there are many improvements!
  - zCX Performance APAR **OA58296** (2/19/2020) provides significant scaling and zIIP eligibility improvements by dramatically reducing switching from zIIPs to GCPs:
    - The more virtual CPUs (VCPs), the greater the benefit
    - Nearly eliminated context switches from zIIPs to GCPs
      - This saved path length and overall CPU
    - Reduced internal latch contention
    - Internal measurements with 16 VCPs showed up to a 50% ETR improvement, double digit ITR improvements, and much smoother scaling

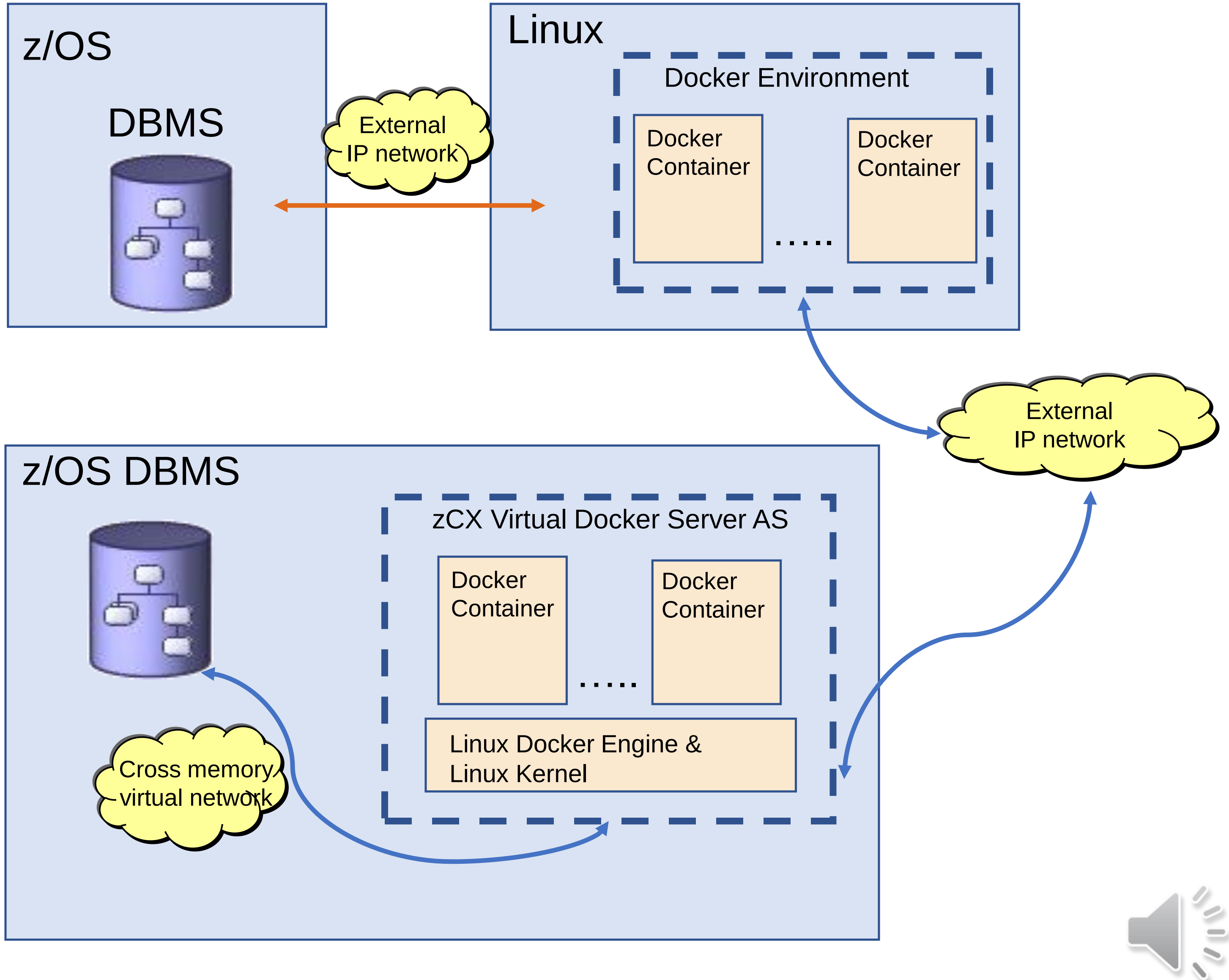
  - zCX Performance APAR **OA59111** (7/1/2020)
    - Provides SIMD (vector) instruction support to the Linux guest and docker containers

  - zCX Performance APAR **OA59573** (9/30)
    - Provides 2G and 1M fixed page support
    - Small hypervisor management improvements
    - See OA59573 charts for performance benefits

zCX Performance Considerations: Network Configuration

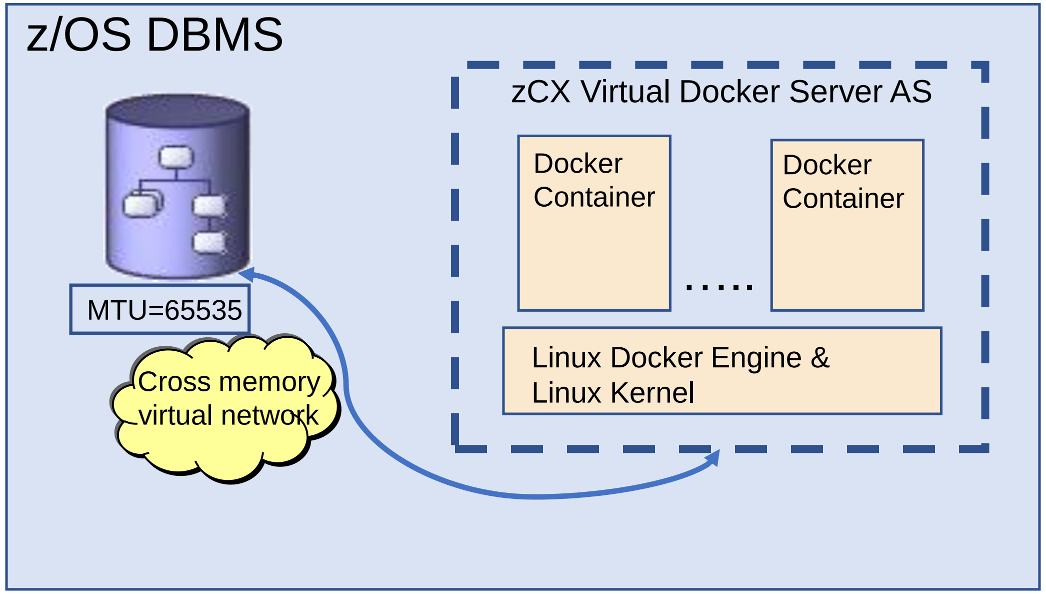
IBM zCX – Moving Docker containers to zCX

- Application tier running in Docker Container on Linux server
  - All communication with Data tier must traverse external network
- Application tier running in Docker container within zCX
  - Co-locating Application tier with Data tier can significantly reduce network latency
  - Reduced network latency for interactive workloads by 45% while increasing network transaction rates by 81%
  - Reduced network latency for streaming workloads by 67% while increasing throughput by over 200%



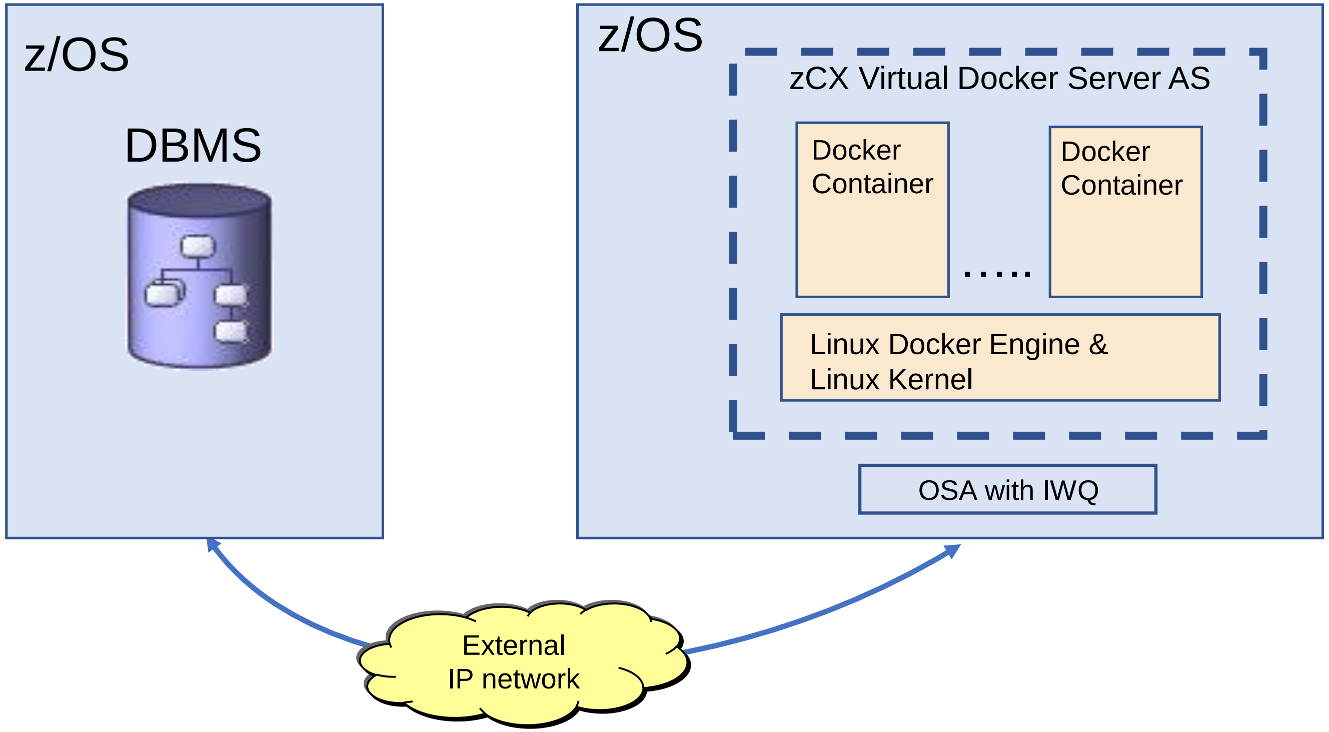
IBM zCX – Optimizing cross memory virtual network

- Virtual network not constrained to packet size limits imposed on physical networks
- When streaming data between the Application and Data tiers, using a larger MTU can provide significant benefits
  - Reduced network latency by 44% while increasing throughput by 80%
  - Reduced network related costs on GCPs by 34% and by 60% on zIIPs


z/OS DBMS
MTU=65535
Cross memory virtual network
zCX Virtual Docker Server AS
Docker Container
.....
Docker Container
Linux Docker Engine & Linux Kernel

IBM zCX – Considerations for non co-located zCX

- Application tier and Data tier running in different z/OS LPARs
  - All communication with Data tier must traverse external network
- Configure Inbound Workload Queuing (IWQ) on OSA-Express
  - Better preserve order of packets delivered to zCX and utilize zIIPs for more network processing
  - Reduced network latency for interactive workloads by 26% while improving network transaction rates by 34%
  - Move nearly 40% of network processing for interactive workloads to zIIPs

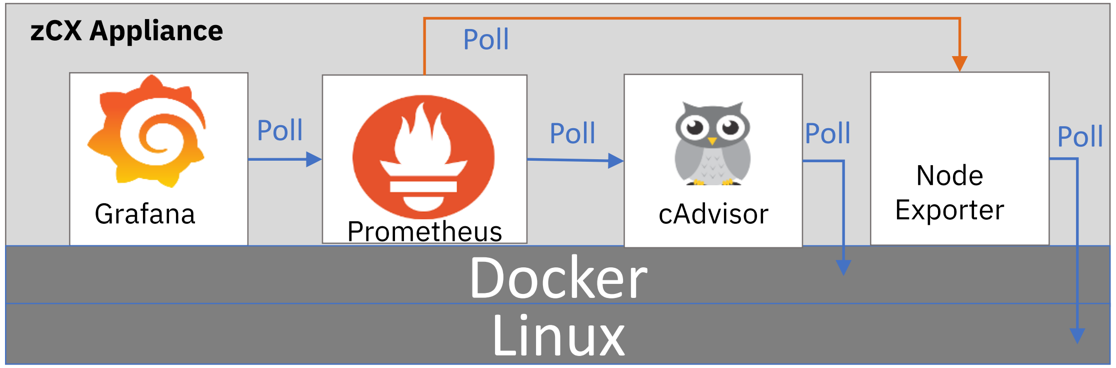


IBM zCX – Communications Server Latest Service

- Make sure you have the latest service before you start as there are many improvements!
  - Communications Server zCX APARs **PH16581** and **OA58300** (11/27/19)
    - Enhancements to support Inbound Workload Queueing (IWQ) for zCX workloads using OSA-Express in QDIO mode
    - Significant offload of zCX network processing to zIIPs
    - Improvements in blocking/batching of work elements for more efficient processing zCX traffic

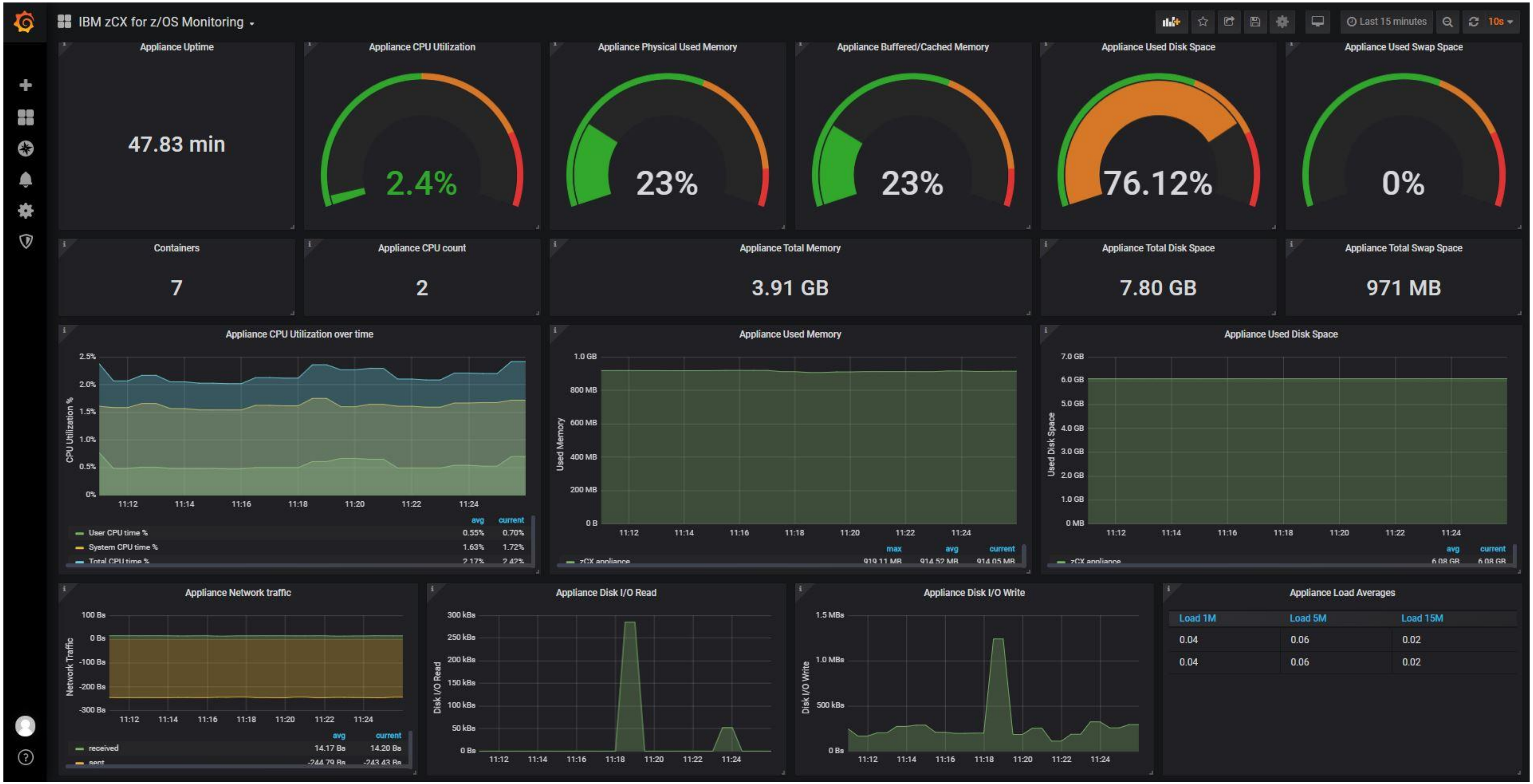
Docker Performance Monitoring

Appliance and Container Monitoring



- zCX provides a sample Grafana dashboard for monitoring the Appliance and Docker containers
- Each of these four components runs in a separate container and play the following roles:
  - Node-Exporter exposes metrics about the Linux operating system.
  - cAdvisor exposes metrics about containers.
  - Prometheus collects the data of the preceding components.
  - Grafana visualizes data that it pulls from Prometheus.
- See this link for up-to-date instructions to run these monitoring tools https://github.com/ambitus/linux-containers/tree/master/examples/grafana
- zCX Grafana dashboard template can be obtained from https://grafana.com/grafana/dashboards/11855
- See the zCX Redbook: Getting Started with z/OS Container Extensions and Docker: sg248457.pdf
  - Has Instructions on how to:
    - Install the required monitoring parts
    - Obtain and adjust the zCX sample dashboards
    - Create your own dashboards
  - But the links with instructions and zCX Grafana dashboard are more current and should be used

zCX Sample: A view of zCX Appliance Level Metrics



Appliance Uptime: 47.83 min

Appliance CPU Utilization: 2.4%

Appliance Physical Used Memory: 23%

Appliance Buffered/Cached Memory: 23%

Appliance Used Disk Space: 76.12%

Appliance Used Swap Space: 0%

Containers: 7

Appliance CPU count: 2

Appliance Total Memory: 3.91 GB

Appliance Total Disk Space: 7.80 GB

Appliance Total Swap Space: 971 MB

	avg	current
User CPU time %	0.55%	0.70%
System CPU time %	1.63%	1.72%
Total CPU time %	2.17%	2.42%

	max	avg	current
zCX appliance	919.11 MB	914.52 MB	914.05 MB

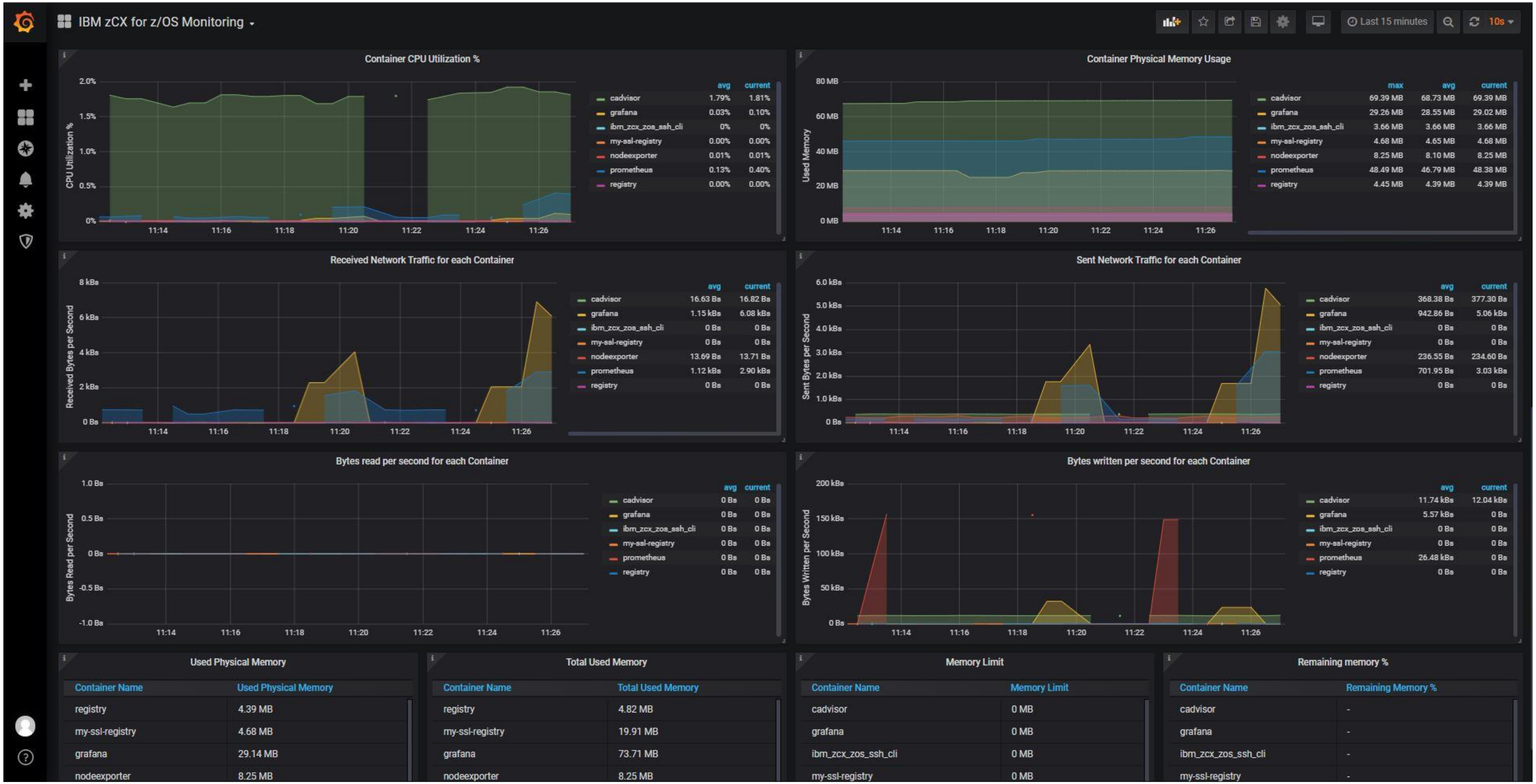
	avg	current
zCX appliance	6.08 GB	6.08 GB

	avg	current
received	14.17 Bs	14.20 Bs
sent	-244.79 Bs	-243.43 Bs

Appliance Load Averages

Load 1M	Load 5M	Load 15M
0.04	0.06	0.02
0.04	0.06	0.02

zCX Sample: A view of Container Level Metrics



IBM zCX for z/OS Monitoring

Container CPU Utilization %

	avg	current
cadvisor	1.79%	1.81%
grafana	0.03%	0.10%
ibm_zcx_zos_ssh_cli	0%	0%
my-ssl-registry	0.00%	0.00%
nodeexporter	0.01%	0.01%
prometheus	0.13%	0.40%
registry	0.00%	0.00%

Container Physical Memory Usage

	max	avg	current
cadvisor	69.39 MB	68.73 MB	69.39 MB
grafana	29.26 MB	28.55 MB	29.02 MB
ibm_zcx_zos_ssh_cli	3.66 MB	3.66 MB	3.66 MB
my-ssl-registry	4.68 MB	4.65 MB	4.68 MB
nodeexporter	8.25 MB	8.10 MB	8.25 MB
prometheus	48.49 MB	46.79 MB	48.38 MB
registry	4.45 MB	4.39 MB	4.39 MB

Received Network Traffic for each Container

	avg	current
cadvisor	16.63 Bs	16.82 Bs
grafana	1.15 kBs	6.08 kBs
ibm_zcx_zos_ssh_cli	0 Bs	0 Bs
my-ssl-registry	0 Bs	0 Bs
nodeexporter	13.69 Bs	13.71 Bs
prometheus	1.12 kBs	2.90 kBs
registry	0 Bs	0 Bs

Sent Network Traffic for each Container

	avg	current
cadvisor	368.38 Bs	377.30 Bs
grafana	942.86 Bs	5.06 kBs
ibm_zcx_zos_ssh_cli	0 Bs	0 Bs
my-ssl-registry	0 Bs	0 Bs
nodeexporter	236.55 Bs	234.60 Bs
prometheus	701.95 Bs	3.03 kBs
registry	0 Bs	0 Bs

Bytes read per second for each Container

	avg	current
cadvisor	0 Bs	0 Bs
grafana	0 Bs	0 Bs
ibm_zcx_zos_ssh_cli	0 Bs	0 Bs
my-ssl-registry	0 Bs	0 Bs
prometheus	0 Bs	0 Bs
registry	0 Bs	0 Bs

Bytes written per second for each Container

	avg	current
cadvisor	11.74 kBs	12.04 kBs
grafana	5.57 kBs	0 Bs
ibm_zcx_zos_ssh_cli	0 Bs	0 Bs
my-ssl-registry	0 Bs	0 Bs
prometheus	26.48 kBs	0 Bs
registry	0 Bs	0 Bs

Used Physical Memory

Container Name	Used Physical Memory
registry	4.39 MB
my-ssl-registry	4.68 MB
grafana	29.14 MB
nodeexporter	8.25 MB

Total Used Memory

Container Name	Total Used Memory
registry	4.82 MB
my-ssl-registry	19.91 MB
grafana	73.71 MB
nodeexporter	8.25 MB

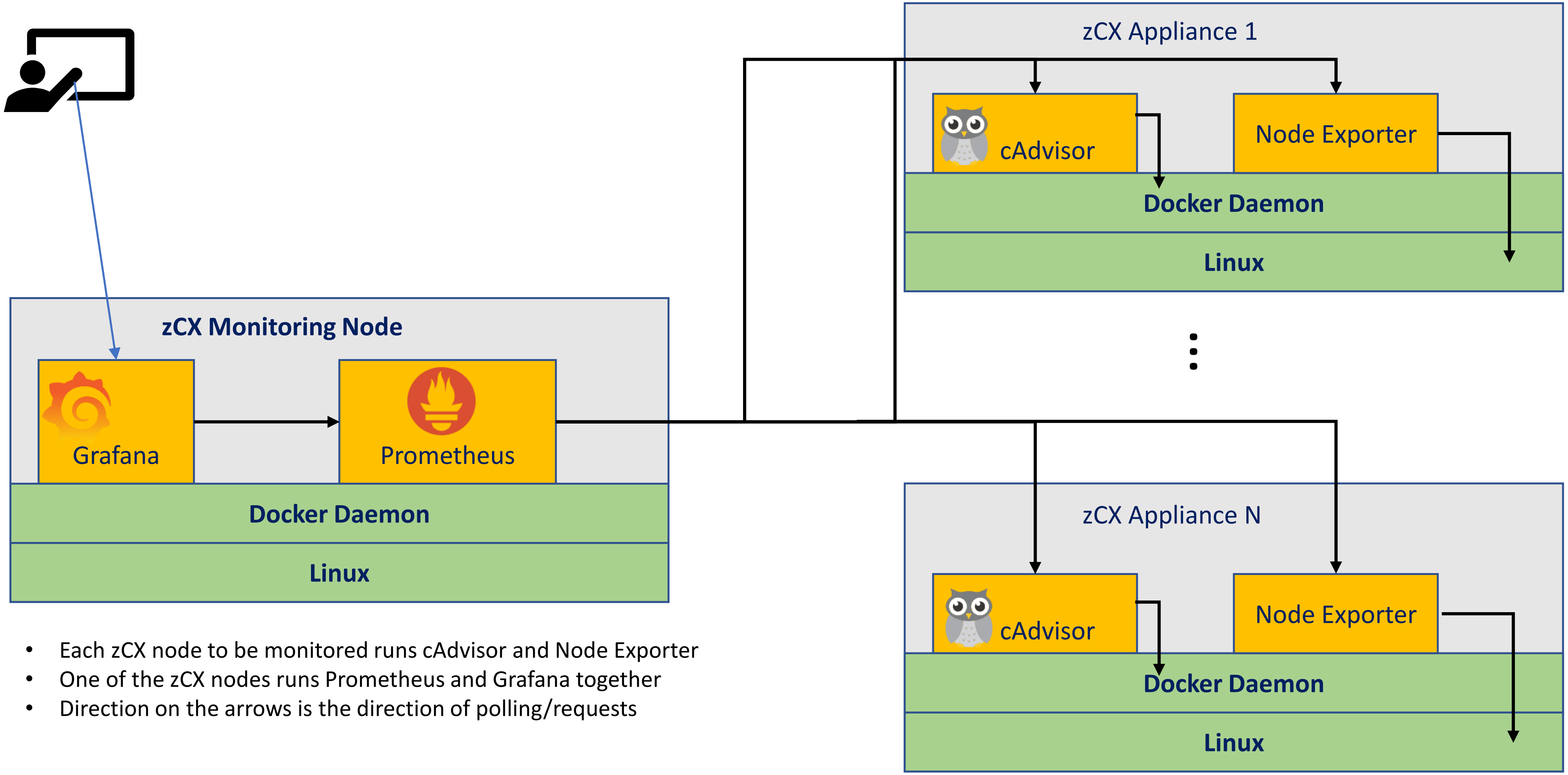
Memory Limit

Container Name	Memory Limit
cadvisor	0 MB
grafana	0 MB
ibm_zcx_zos_ssh_cli	0 MB
my-ssl-registry	0 MB

Remaining memory %

Container Name	Remaining Memory %
cadvisor	-
grafana	-
ibm_zcx_zos_ssh_cli	-
my-ssl-registry	-

A model for monitoring multiple zCX appliances



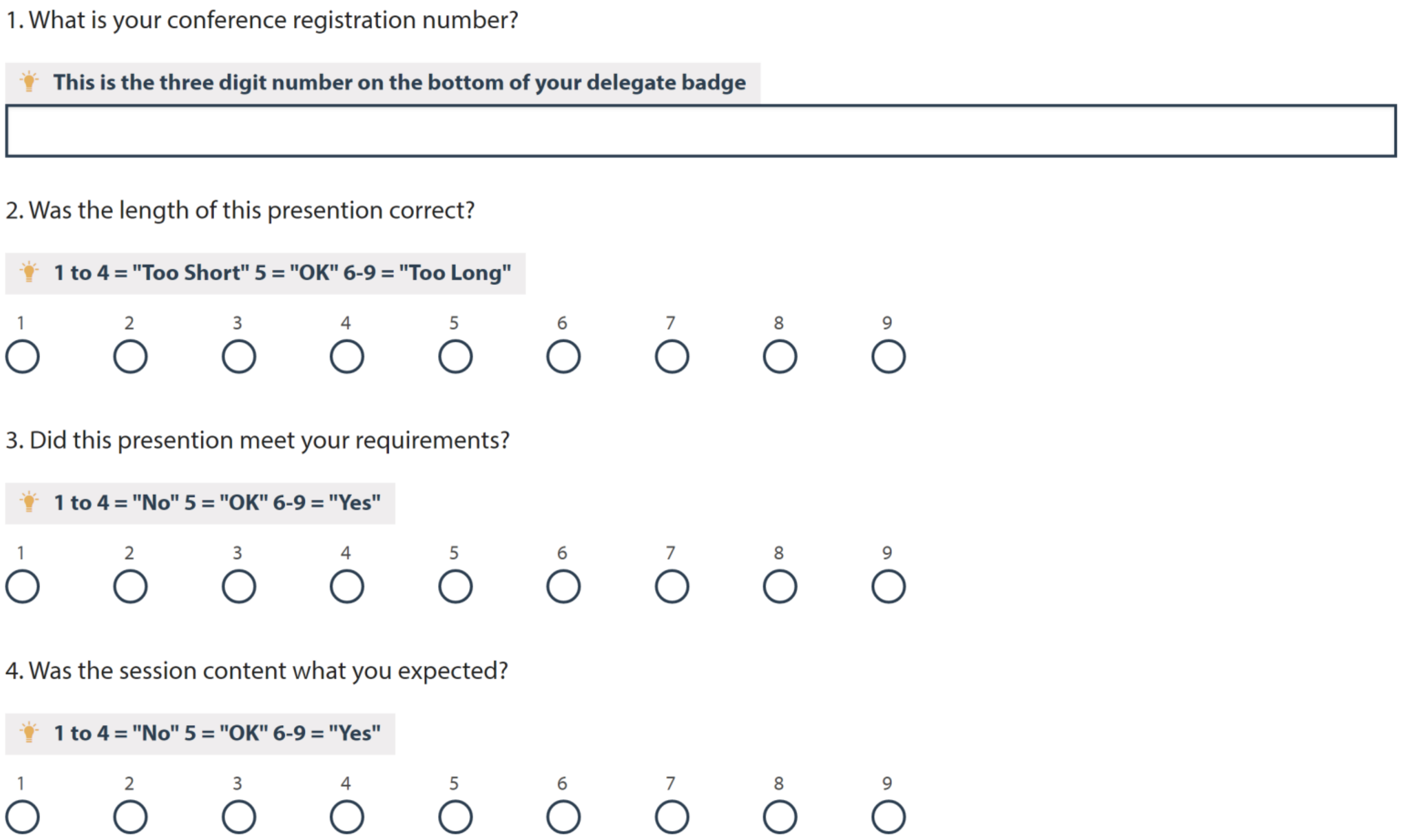
- Each zCX node to be monitored runs cAdvisor and Node Exporter
- One of the zCX nodes runs Prometheus and Grafana together
- Direction on the arrows is the direction of polling/requests

zCX – Reference Material

- z/OS V2R4 Communications Server Performance Summary Report
- White Paper: Ready for the Cloud with IBM z/OS Container Extensions by IBM IT Economics Consulting & Research
- zCX Documentation
- zCX Redbook: Getting Started with z/OS Container Extensions and Docker
- z/OS V2R4 MVS Planning: Workload Management
- Hot topics articles:
  - Rapid Containers: Improving zCX Runtime Performance
  - Running Linux on IBM Z Docker Containers Inside z/OS

- Do it online at http://conferences.gse.org.uk/2020/feedback/1AE

- This session is 1AE



1. What is your conference registration number?

This is the three digit number on the bottom of your delegate badge

2. Was the length of this presention correct?

1 to 4 = "Too Short" 5 = "OK" 6-9 = "Too Long"

1	2	3	4	5	6	7	8	9
○	○	○	○	○	○	○	○	○

3. Did this presention meet your requirements?

1 to 4 = "No" 5 = "OK" 6-9 = "Yes"

1	2	3	4	5	6	7	8	9
○	○	○	○	○	○	○	○	○

4. Was the session content what you expected?

1 to 4 = "No" 5 = "OK" 6-9 = "Yes"

1	2	3	4	5	6	7	8	9
○	○	○	○	○	○	○	○	○

GSE UK Conference 2020 Charity

- The GSE UK Region team hope that you find this presentation and others that follow useful and help to expand your knowledge of z Systems.
- Please consider showing your appreciation by kindly donating a small sum to our charity this year, NHS Charities Together. Follow the link below or scan the QR Code:

http://uk.virginmoneygiving.com/GuideShareEuropeUKRegion