IDUG

Leading the Db2 User Community for 35 Years

IDUG

2023 EMEA Db2 Tech Conference

Db2 purescale on z/Linux

Dale McInnis

IBM Canada Ltd.

Session Code: D3(9005)

Platform: Linux, UNIX and Windows

#IDUGDb2

Who am I

- Certified Technical Specialist (level 3) on the America's Data and AI Technical Sales team
- 35 years with IBM, 25 years as the Db2 LUW High Availability Architect
- Vice Chair for Canadian Information Processing Society – Ontario Canada
  - Information Systems Professional (I.S.P) Certified
  - Information Technology Certified Professional (ITCP)
- IBM WorldWide Performance and Availability Community co-leader
- IBM Academy of Technology Member
- IBM Academy PREVAIL Conference Lead
- 11 Patents in database resiliency
- Published author
- IDUG Speaker Hall of Fame member
- PhD Student at Ontario Technical University

Db2 pureScale on LinuxONE

Db2 on Z/LinuxONE support

Olqx{RQH Hp shuruLL

IBM z14 (M/T 3906)
- Announced 7/17/2017
- 5.2 GHz
- Up to 170 cfg. cores
- CP, IFL, ICF, zIIP
- Up to 32 TB cfg. Memory

IBM z14 ZR1 (M/T 3907)
- Announced 4/10/2018
- 4.5 GHz
- 19" Rack
- Up to 30 cfg cores (6 CP)
- CP, IFL, ICF, zIIP
- Up to 8 TB cfg. Memory

Olqx{RQH LLL

IBM z15 T01 (M/T 8561)
- Announced 09/12/2019
- 5.2 GHz
- Up to 190 cfg. cores
- CP, IFL, ICF, zIIP
- Up to 40 TB cfg. Memory

IBM z15 T02 (M/T 8562)
Announced 04/14/2020
4.5 GHz
Up to 65 cfg. cores
CP, IFL, ICF, zIIP
Up to 16 TB cfg. Memory

Olqx{RQH Hp shuru7

IBM z16 A01 (M/T 3931)
- Announced April 5th, 2022
- 5.2 GHz
- Up to 200 cfg. cores
- CP, IFL, ICF, zIIP
- Up to 40 TB cfg. Memory

IBM z16 Single Frame Model A02, IBM z16 Rack Mount - (M/T 3932)
- Announced April 4th, 2023
- 4.6 GHz
- Up to 68 cfg. cores
- CP, IFL, ICF, zIIP
- Up to 16 TB cfg. Memory

Integrated Facility for Linux (IFL) on IBM z16

Run Linux workloads with a dedicated, fast, and cost-effective processor

Dedicated and cost-effective

An IFL is a dedicated processor for Linux® workloads, attractively priced, and supporting Linux for IBM zSystems, IBM z/VM®, KVM, and Red Hat® OpenShift® deployed in virtual machines

IFLs allow for a very high virtual Linux server density resulting in reduced expenses in the areas of operational efforts, software licensing, floor space, and energy consumption.

An IFL does not increase charges for IBM zSystems software running on "standard" processors, nor does it affect the MSU rating or the IBM zSystems model designation

Massive processing capacity

An IFL is always a full capacity processor, independent of the capacity of the other processors on the server

Up to 200 user-configurable IFLs running at 5.2 GHz and up to 40 TB Redundant Array of Independent Memory on an IBM z16™ A01

IFLs can be simply added into the existing server, on the fly, without affecting the business

Manageable by PR/SM™ or IBM Dynamic Partition Manager in LPAR with dedicated or shared processors

Designed for performance

On-chip acceleration provides faster AI insights at scale, increases data transfer rates and rapid encryption

IFL supports enhanced Simultaneous Multi-Threading (SMT) and Single Instruction Multiple Data (SIMD) technologies to improve performance

IBM HiperSockets™, Shared Memory Communication (SMC), and zcsfs file system functionality provide high speed communication between logical partitions

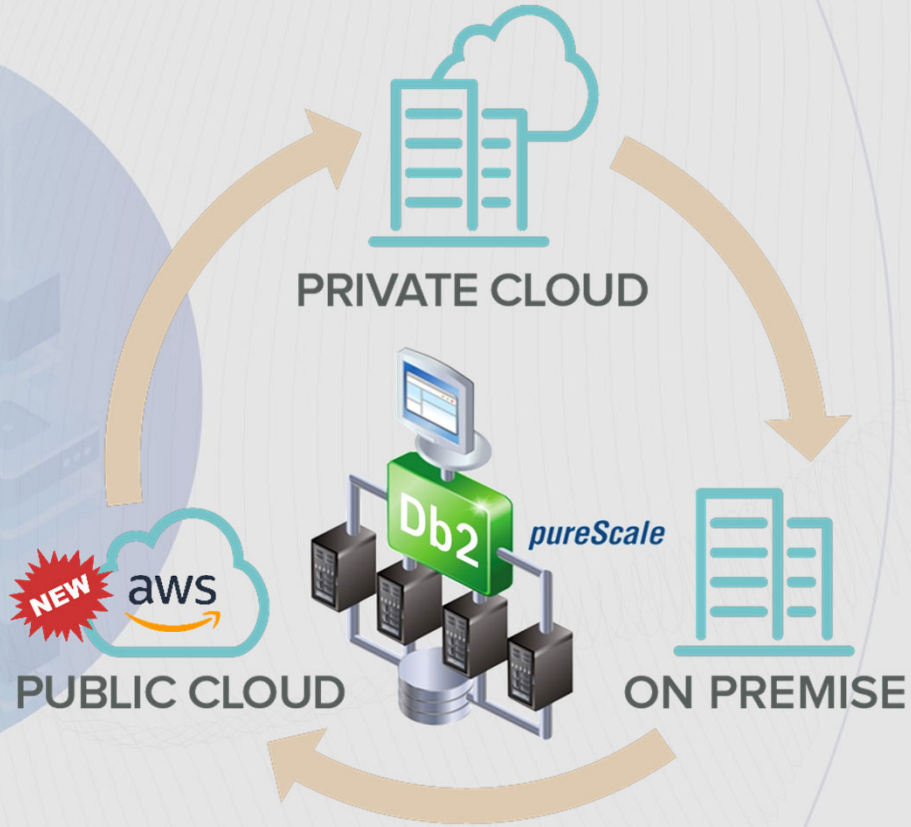
IFL technology supports On/Off Capacity on Demand, Capacity Upgrade on Demand for the non-disruptive addition, and Capacity BackUp for emergency situations

1 IFL can be do as much work as 1 X86 to 33 X86 cores Depending on the workload

I use a 10:1 ratio as a estimate

Link to learn more

Db2 LUW's five 9s solution is *pureScale*



pureScale Heritage, Evolution, and Core Values

pureScale has

decades of technological maturity

- Db2 for LUW first introduced in 1993
- pureScale technology first introduced in 1990 on zOS (Parallel Sysplex)
- Sysplex ported to Db2 LUW as pureScale in 2009

rich enterprise functionality

- SQL support - leverage existing (or widely available) skills
- Fully consistent ACID – including multi-statement transactions
- Fine grained access control – restrict access by row/column

pureScale runs on

2009
- POWER AIX with Infiniband (IB)

2012 Jan
- POWER AIX with RoCE

2012 Dec
- x86 RHEL and SLES with IB and RoCE

2014
- AIX + Linux with TCP/IP support

2017
- POWER Linux

2021
- Linux on IBM Z

2022 June NEW
- Self-Managed service on AWS Marketplace

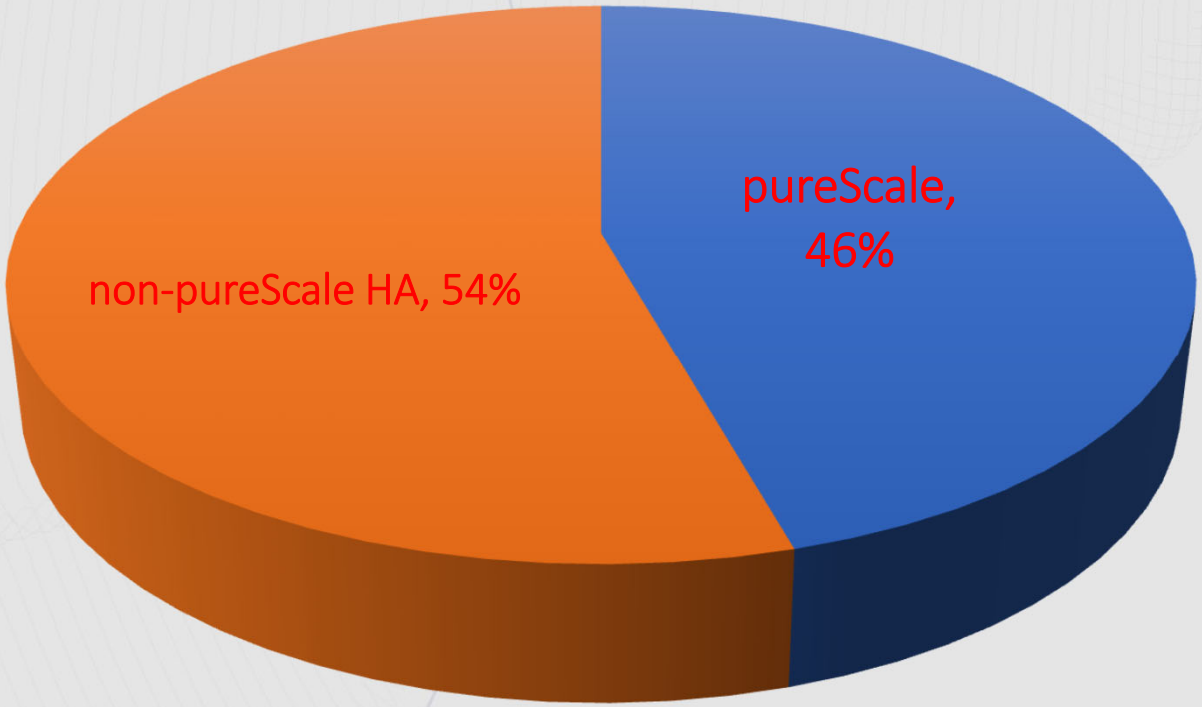
pureScale is

battled tested in the world's most demanding environments …

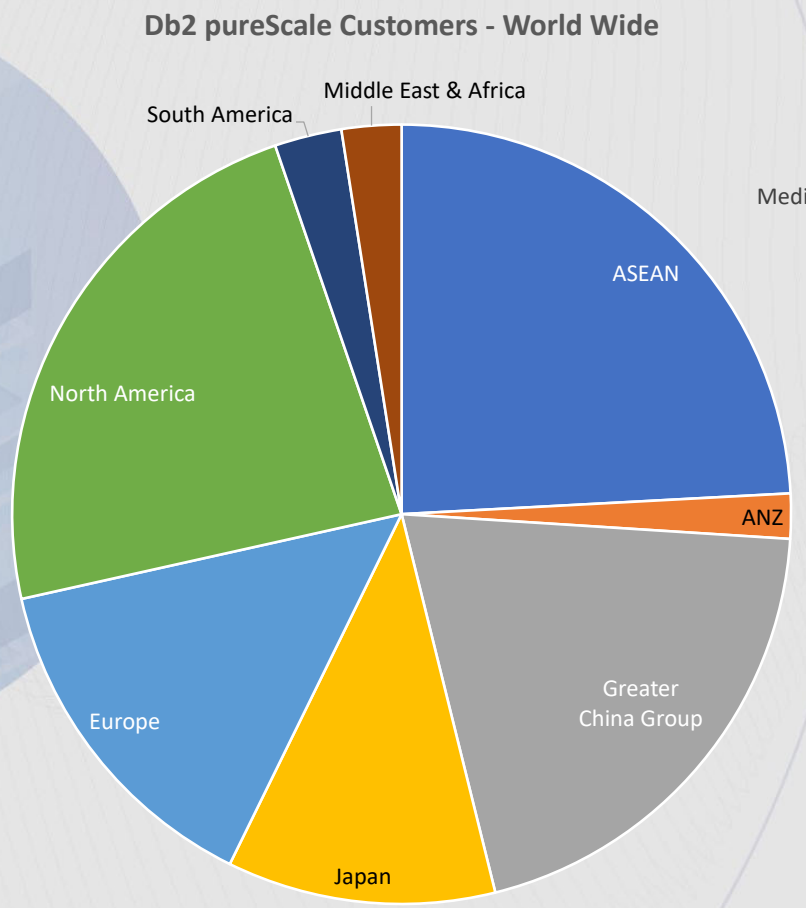
7

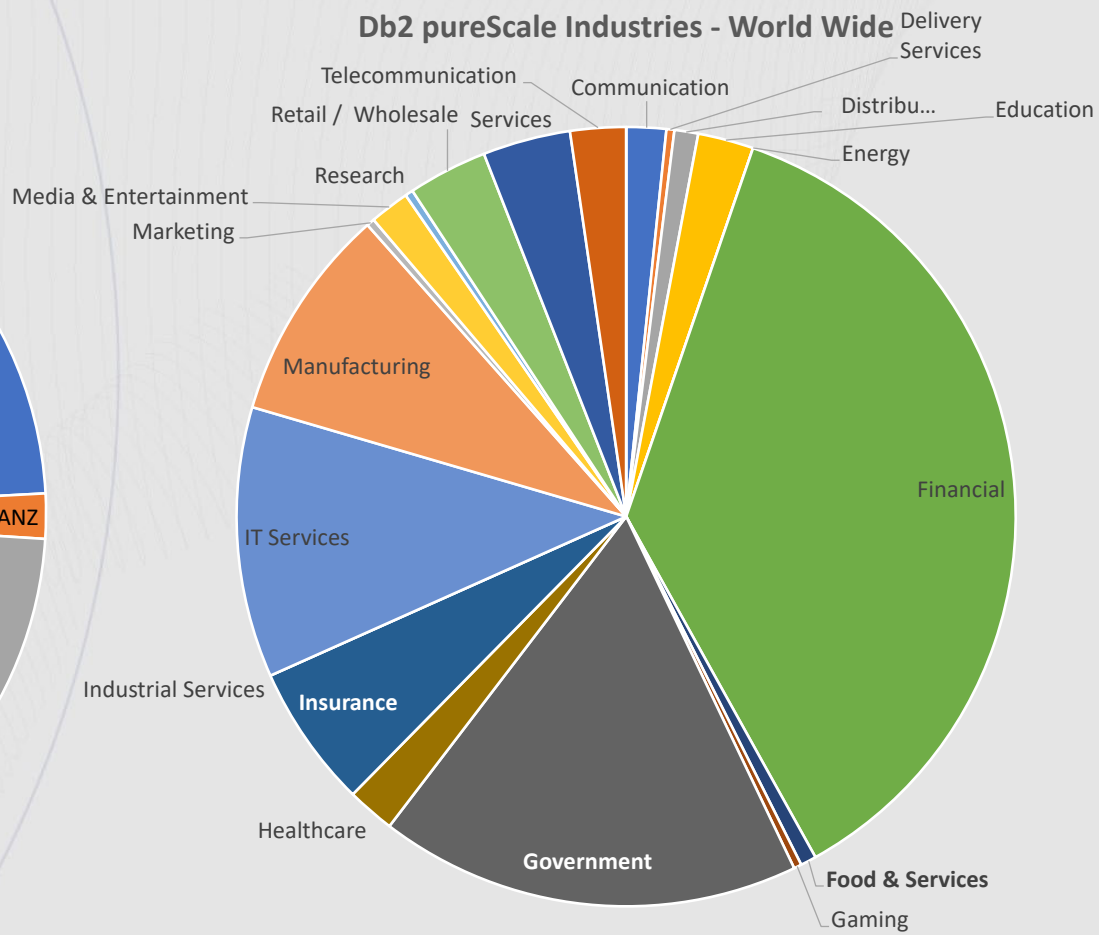
PureScale Adoption for High Availability

Db2 pureScale Vs non-pureScale HA Systems World Wide

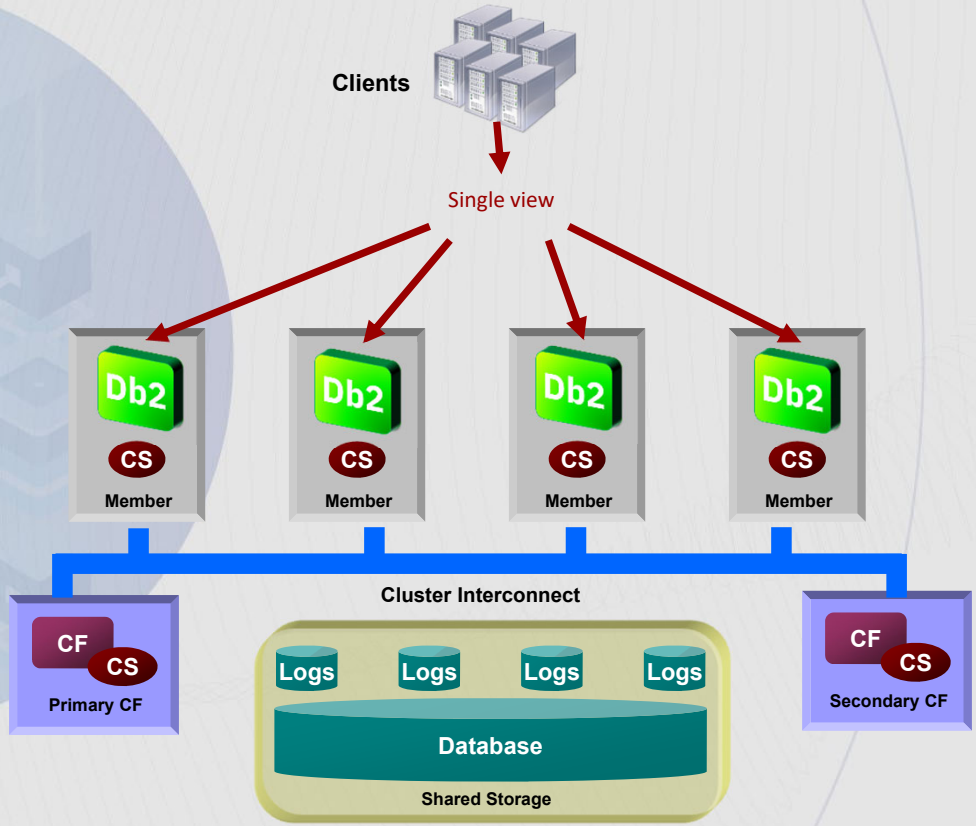


pureScale Successes at Every Geography and Every Industry





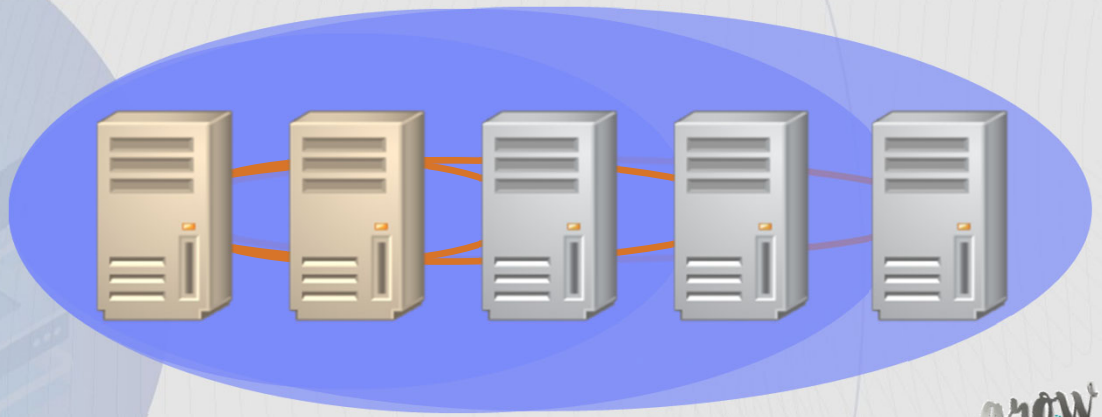
Features Overview

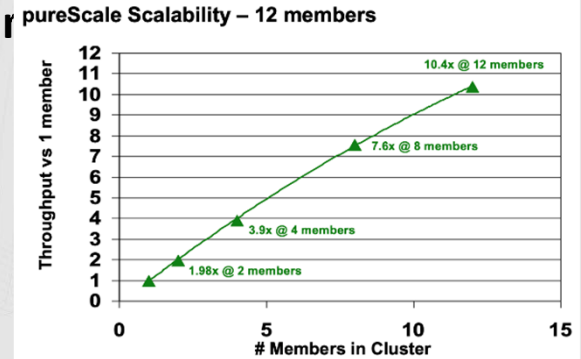


- Automatic Workload Balancing
- Cluster of Db2 nodes running on bare-metals or VMs (POWER, x86, Z, AWS EC2)
- CF — Caching Facility (CF) - Centralized global lock and page cache management
- CS — Cluster Services - Integrated failure detection and recovery automation
- High speed interconnect (RDMA & TCP/IP)
- Shared data

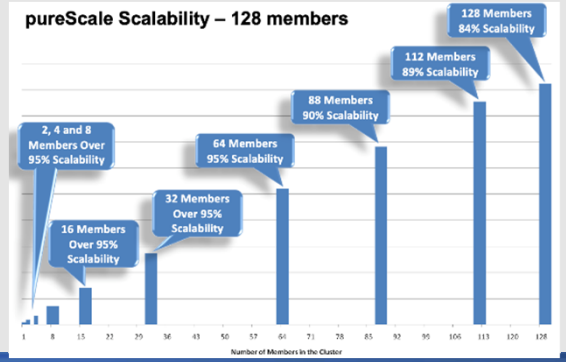
Extreme Capacity and Application Transparency

- **Db2 pureScale is designed to grow to whatever capacity your busir**





- **Take advantage of extra capacity instantly**
  - No need to modify your application code
  - No need to tune your database infrastructure



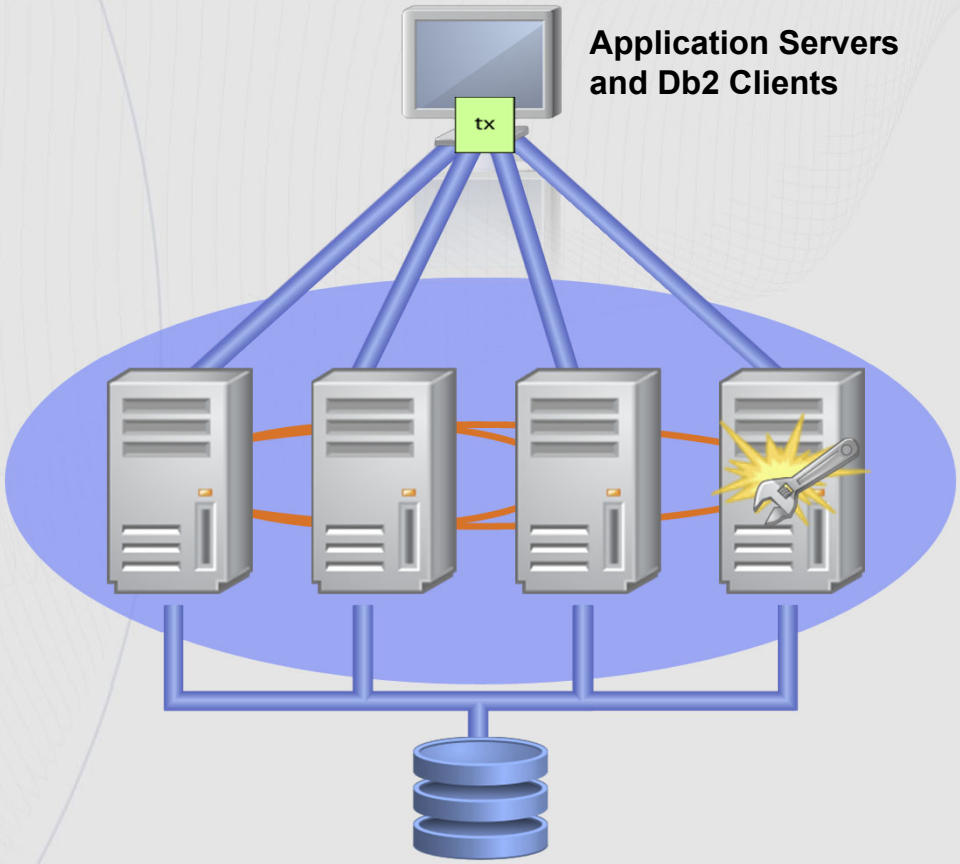
Architecture scales to 100+ nodes

Your DBAs can add capacity without re-tuning or re-testing

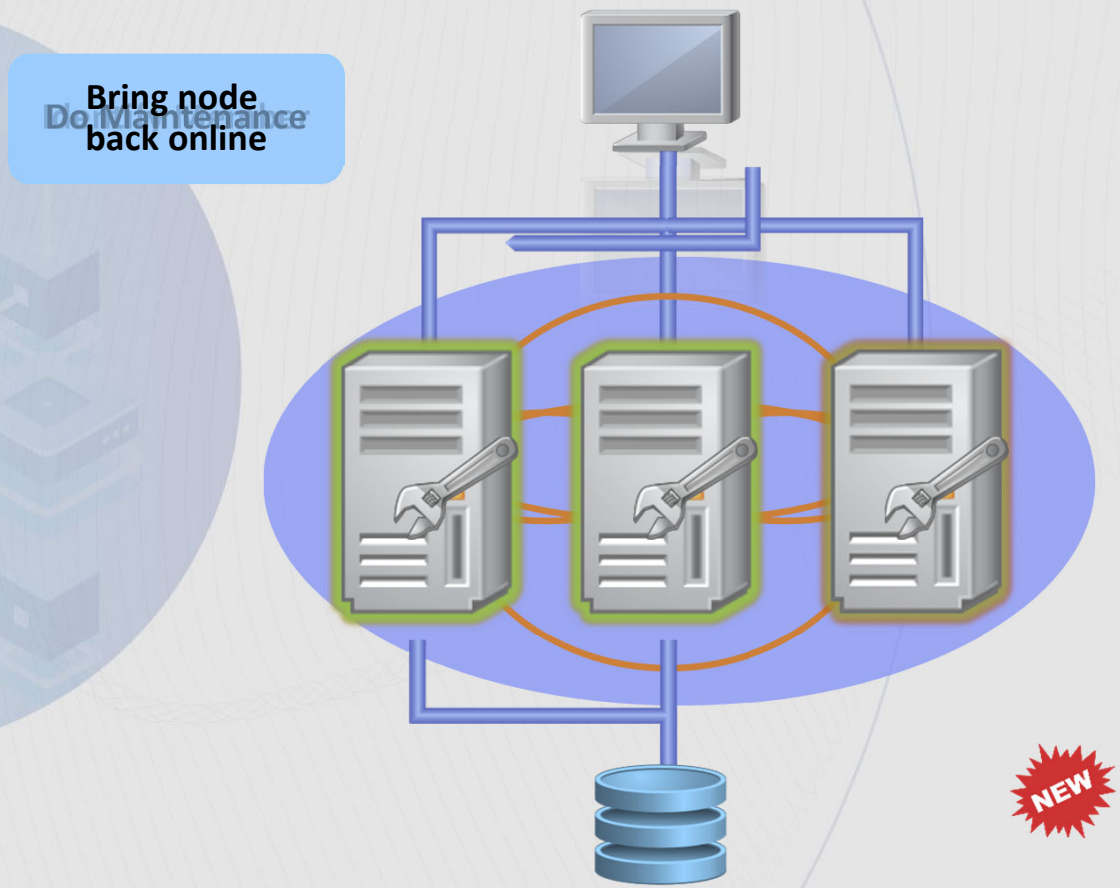
Your developers don't even need to know more nodes are being added

Recover Instantaneously From Node Failure (*Unplanned* Outages)

- **Protect from infrastructure related outages**
  - Redistribute workload to surviving nodes immediately
  - Completely redundant architecture
  - Recover in-flight transactions on failing node in seconds (including detection of the problem)



Minimize the Impact of *Planned* Outages



- **Keep your cluster up during various maintenance operations:**
  - Db2 mod/fix pack maintenance – "Rolling Update"
  - OS fixes
  - Hardware updates
  - Administration

- **Supports multiple concurrent hosts maintenance** *(as long as quorum is maintained)*

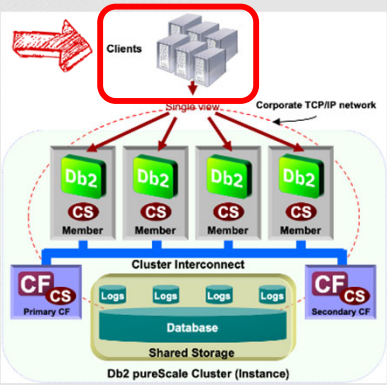
NEW

NEW

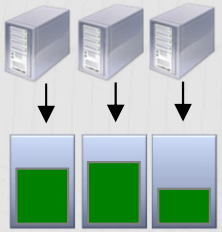
- **Smart and Fast fix pack or mod pack update**
  - Single command on any host to update a cluster of ANY size (maximum concurrency guaranteed!)

pureScale Clients

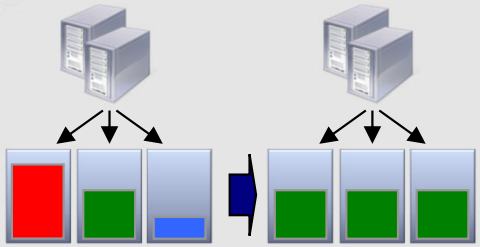
- Db2 front end for customers to submit work (queries, updates, etc.) to server via command line processor (CLP), Java/C apps, etc.



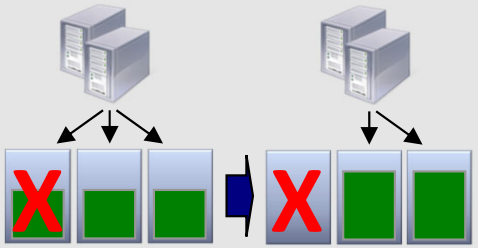
- **Q: Which member(s) will my application run on ?**
  - **Client Affinity**
    - Direct different groups of clients or workloads to specific members
    - Consolidate separate workloads/applications on same database infrastructure
    - Define list of members for failover purposes



  - **Workload Balancing (WLB)**
    - Application requests are balanced across all members or subsets of members
    - Takes server load of members into consideration

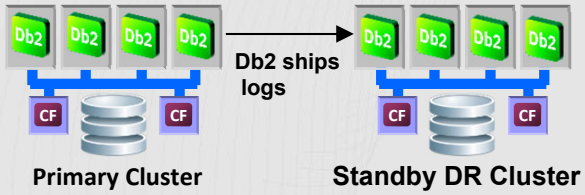


  - **Automatic Client Reroute (ACR)**
    - Client automatically connected to healthy member in case of member failure
    - May be seamless in that no error messages returned to client
    - Application may have to re-execute the transaction

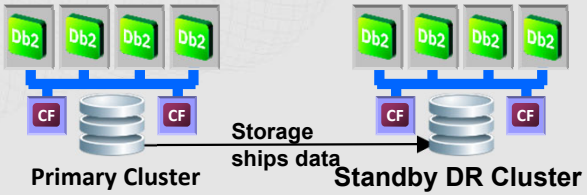


Disaster Recovery (DR) Options

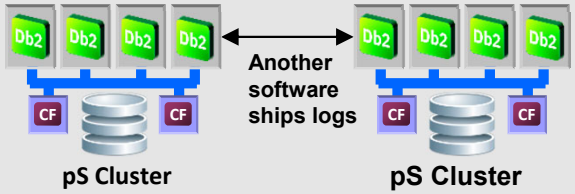
- 1. HADR
  - Db2 is responsible to ship data across
  - Replication between a database in a primary pS cluster and a standby pS cluster with a matching member topology
  - Active/passive DR



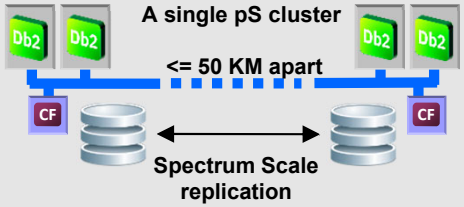
- 2. Storage Replication
  - The storage layer is responsible to ship data across
  - Synchronous or asynchronous disk-based replication between a database in a primary pS cluster and a standby pS cluster with a matching member topology
  - Active/passive DR



- 3. Q Replication / InfoSphere Change Data Capture (CDC)
  - Another software is responsible to ship data across
  - Logical replication between two databases
  - Can be active/active DR



- 4. Geographically Dispersed pureScale Cluster (GDPC)
  - Data is replicated on each I/O synchronously
  - Single pS cluster "stretched" over two sites with half of members/CFs at each site
  - Active/active DR



Continuous Investments & Commitment to pureScale over the years

Security

V11.5.0.0

- Host-based firewall enablement is now an option for single-node, DPF, and pureScale Db2 configurations, via the new Db2_FIREWALL_PORT_RANGE registry variable.

Resiliency & Continuous Availability

V11.1.3.3

- Online add/drop CF
- Concurrent hosts maintenance

V11.1.4.4

- Enhanced multi-tenancy HA by supporting maintenance at database mount resource level.
- Support Cluster Configuration Repository (CCR) in Spectrum Scale for better HA

V11.5.8.0

- Lock timeout avoidance with members in EHL mode
- Priority Flow Control (PFC) support on AIX

Recovery Performance

V11.1.3.3

- Faster crash recovery with parallel cleanup in CF global buffer pool and global lock manager

V11.5.0.0

- Faster CF recovery on Ethernet failure scenario

Monitoring & Problem Determination

V11.1.3.3

- Point-to-point RDMA ping

V11.1.4.4

- New db2pd option to provide CF insight
- Additional CF commands tracking through MON_GET_CF_CMD table function
- Cluster manager repair improvement

V11.5.5.0

- Lightning fast RDMA ping validation
- Periodic Automatic cluster topology validation

V11.5.8.0

- Periodic Automatic RDMA adapter liveliness monitor

Usability & Performance

V11.1.3.3

- "Click & Go" performance data collection
- Concurrent write access with index creation
- Multiple cross-invalidation support in TCP/IP

V11.1.4.4

- Reduce table space size via extent movement

V11.5.0.0

- Improved re-use of free pages across members
- Faster LOAD for range-partitioned tables
- Enhanced page cleaning in CF for busy clusters
- Multiple cross-invalidation support in RDMA
- Currently committed semantics across members

V11.5.6.0

- Default split diagnostic path to reduce contention

V11.5.8.0

- Parallel page registration in CF led upto 10x faster in large bufferpool page read & write

Up & Running

V11.1.4.4

- Faster cluster setup with simplified network monitoring configuration and requirements

V11.5.0.0

- GDPC: Lower Total cost of ownership(TCO) via automatic configuration of public network resources

V11.5.5.0

- Concurrent online fix/mod pack update

V11.5.8.0

- Smarter concurrent online fix/mod pack update
- Replace uDAPL with Verbs on RHEL 8.x RDMA
- pS with TCP/IP network on AWS marketplace *(already available 2Q 2022)*

Going Forward: Continuous Investments & Commitment to all our customers

- **Covers:**
- All form factors: On-premise and Cloud
- All architectures: POWER, Intel, Z

Supported PureScale Architectures for Db2 v 11.5.8

Current detailed system requirements

Operating Systems Supported		
Power series	**Z14 / Z15**	**X86**
AIX 7.2 TL5+	RHEL 8.6 (Z15)	RHEL 7.9
AIX 7.3 TL0 SP1	RHEL 7.7 (Z14) V 11.5.5.1	RHEL 8.4 – 8.6
RHEL 7.9		SLES 12 SP5
RHEL 8.4		SLES 15 SP4

Hypervisors supported	
AIX	**Linux**
IBM Power VM (LPAR, DPAR, Micro-partition)	IBM Power VM (LPAR, DPAR, Micro-partition)
	KVM in SLES 12
	Red Hat KVM delivered in RHEL
	VMWare ESXi 6.x

Roadmap: pureScale

All Deployment Environments

START Fall 2022

Q4 2022 – V11.5.8.0

- RHEL support
- RDMA on RHEL 8.6+
- Direct IO on Z/Linux (default)
- Z15 support

Q4 2023 – V11.5.9

- Pacemaker Tech Preview (run at)
- Performance enhancements
- Z16

In Progress

2024 - VNext

- ***Pacmaker support GA (run at)***
- Open-source cluster manager
- Align with new Db2 major release

In Progress

2024 or beyond

- Replace uDapl with VERBs (AIX + SLES)
- Mellanox CX-6 support
- CF Performance enhancements
- RHEL 9.x

TBD

Fully-Managed pS on public clouds

- Orthogonal to all of the above.

Timeline and Roadmap subjected to change without notice

SICOOB

- SICOOB is the biggest credit union in Latin America. Supporting more than 400 Cooperatives spread through Brazil.
- Over 1000 VMs, their core system is deployed on top of 4 LinuxONE Emperors II with z/VM clusters, on active-active datacenter topology delivering a highly resilient and Scalable environment.
- Integrating z/VM to Vmware vRealize with z/VM Cloud Connector plugin and Ansible automation SICOOB can deploy virtual machines on LinuxONE and Intel boxes using the same dashboard as a single panel to deliver on demand infrastructure.
- With Db2 HADR, SICOOB guarantees zero data loss with a active/passive database schema between sites that is auto-managed in case of a disaster, shifting seamlessly between instances without offering minimal downtime to their operations.
- Through ICP, containers are deployed simultaneously on LinuxONE and intel on premises and on public cloud providers to support peak processing and flexible alternatives in case of a disaster.
- On the opensource vein, SICOOB uses Grafana to monitor in realtime their entire z/VM environment providing historical trends to better manage resource allocation.

Db2 LinuxONE Requirements

https://www.ibm.com/docs/en/db2/11.5?topic=linux-installation-prerequisites-db2-purescale-feature-z

The Db2 pureScale Feature in Db2 11.5.5.1 only is supported on z14-compatible rack mounted server that supports the 10GbE RoCE Express (FC 0412) Ethernet RoCE adapter.

The Db2 pureScale Feature in Db2 11.5.8 and later is supported on z15-compatible rack mounted server that supports the 10GbE RoCE Express (FC 0412) Ethernet RoCE adapter.

Db2 v 11.5.5.1 minimum disk requirements

	Recommended minimum free disk space	Minimum required free disk space
Disk to extract installation	3 GB	3 GB
Installation path	6 GB	6 GB
/tmp directory	5 GB	4 GB
/var directory	5 GB	2 GB
/usr directory	2 GB	512 MB
Instance home directory	5 GB	1.5 GB[1]
Root home directory	300 MB	200 MB

Db2 LinuxONE 10Gb Cable Information

•Refer to the Z documentation for cable options for the corresponding adapter that is being selected. The cables available for the FC 0412 adapter on z15 are listed under the adapter here: IBM z15 Technical Introduction, under section 4.3 *Network Connectivity*.
•The cables available for the FC 0411 adapter on z14 are listed under the adapter here: IBM z14 (3906) Technical Guide, under section 4.7.3 *Network Connectivity.*

LPAR vs z/VM

When deployed on an RoCE network, Db2 members and CFs must be on their own Logical Partition (LPAR) with dedicated RoCE adapters.

Miscellaneous requirements

Other mandatory changes prior to setting up Db2 pureScale

Complete the following tasks before setting up your Db2 pureScale environment:

- Disable the IPv6 network by setting the relevant entries in /etc/sysctl.conf:

net.ipv6.conf.all.disable_ipv6 = 1 ...

For RHEL 7 systems, set the following required kernel values for Spectrum Scale:

- Add the following entry to /etc/zipl.conf:

vmalloc=4096G

Run the following commands as root to make the change effective:

dracut -f zipl -V reboot

For RHEL 8 systems, set the following required kernel value for Spectrum Scale:

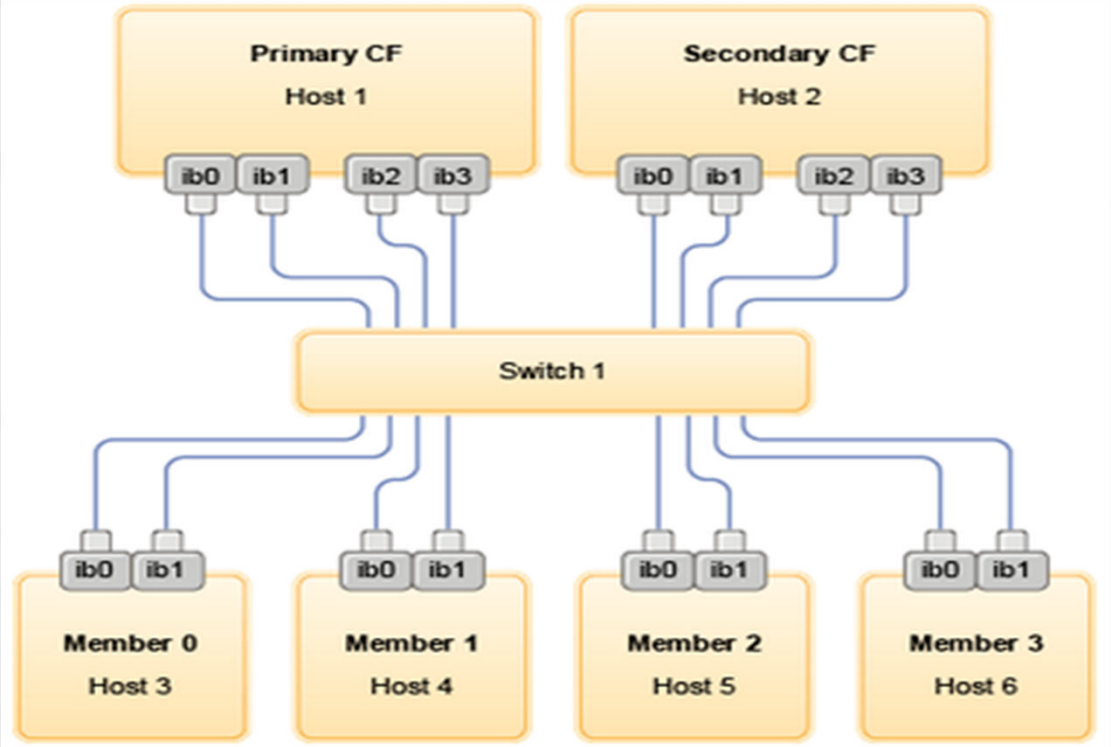
- Use the grubby utility to add vmalloc=4096G for all boot entries on your node:

grubby --update-kernel=ALL --args="vmalloc=4096G" zipl -V reboot

Olqx{ rq LEP]47 vriwz duh uhtxluhp hqw

Linux distribution	Required packages	OpenFabrics Enterprise Distribution (OFED_ package
Red Hat Enterprise Linux (RHEL) 7.7	libibverbs librdmacm rdma-core dapl ibacm libstdc++ glibc gcc-c++ gcc kernel kernel-devel kernel-headers linux-firmware ntp or chrony ntpdate sg3_utils sg3_utils-libs binutils m4 openssh cpp ksh file libgomp make patch perl-Sys-Syslog pam pam-devel lsscj libibumad Infiniband-diags	For RoCE network type run group installation of "InfiniBand Support" package.

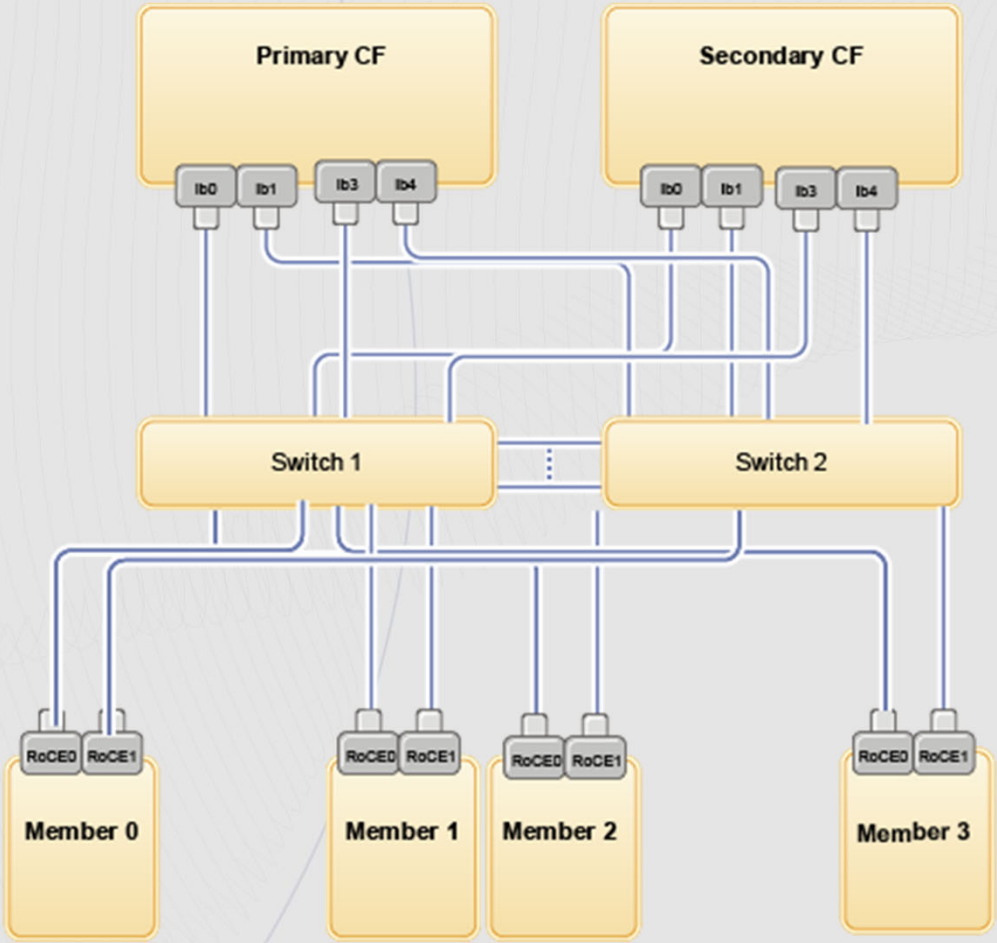
Z14 / Db2 V 11.5.5.1 – only single switch support



Olqx{ rq LEP]48 vriwz duh uhtxluhp hqwv

Linux distribution	Required packages	OpenFabrics Enterprise Distribution (OFED_ package)
Red Hat® Enterprise Linux (RHEL) 8.6	– libibverbs – librdmacm – rdma-core – ibacm – libstdc++ – glibc – gcc-c++ – gcc – kernel – kernel-devel – kernel-headers – linux-firmware – chrony – sg3_utils – sg3_utils-libs – binutils – m4 – openssh – cpp – ksh – file – libgomp – make – patch – perl-Sys-Syslog – pam – libibumad	Install the Red Hat OFED RDMA packages using dnf groupinstall InfiniBand Support.

Z15/ Db2 V 11.5.8 – Dual switch support



PoT Overview

Topology

3.1. System

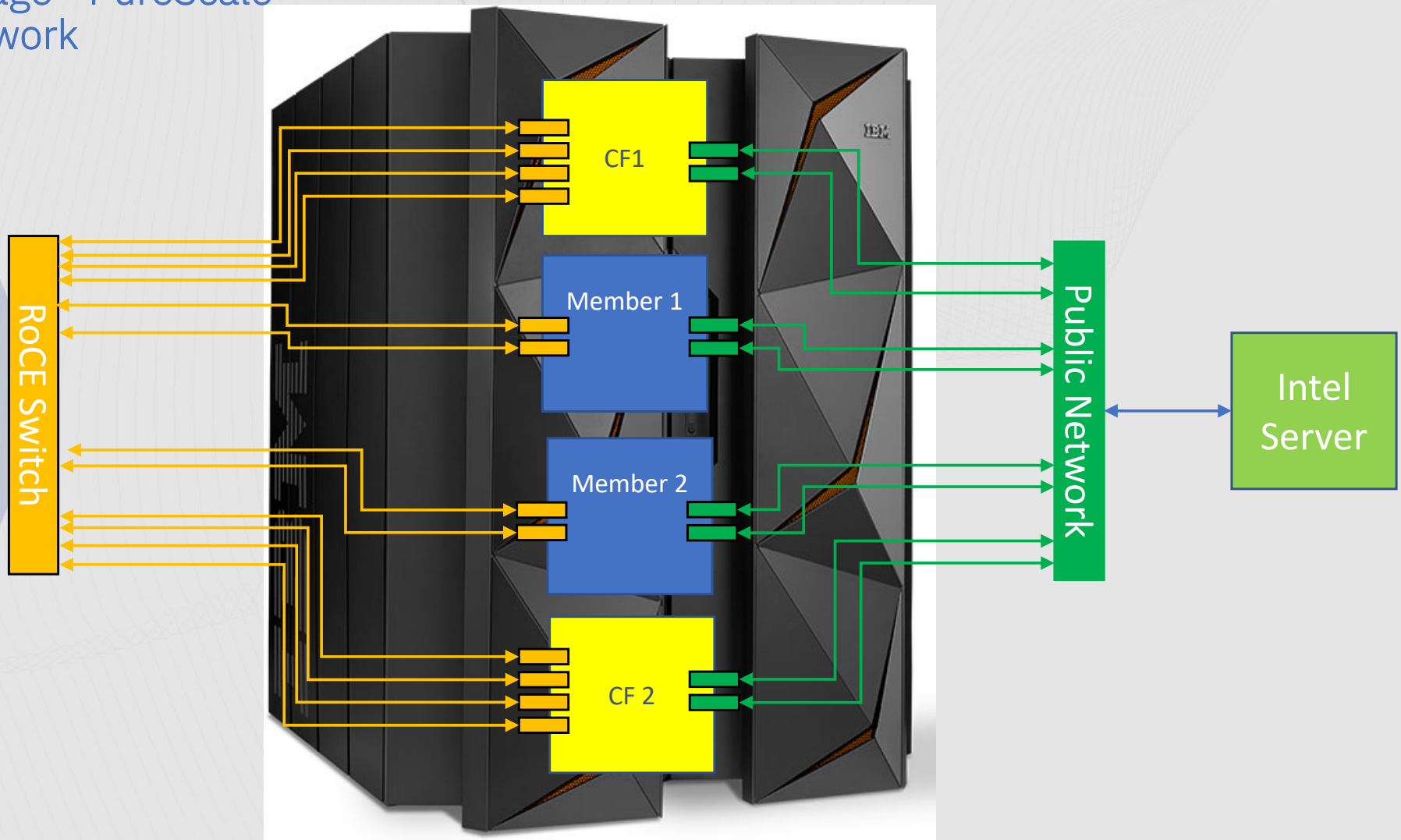
Make: 3906
Model: M05

LPAR	Type	CPU/Type	Memory	Crypto	SMT2
L101	Native	4 IFL (dedicated)	32GB	No	Yes
L102	Native	4 IFL (dedicated)	32GB	No	Yes
L103	Native	8 IFL (shared)	32GB	No	Yes
L104	Native	8 IFL (shared)	32GB	No	Yes

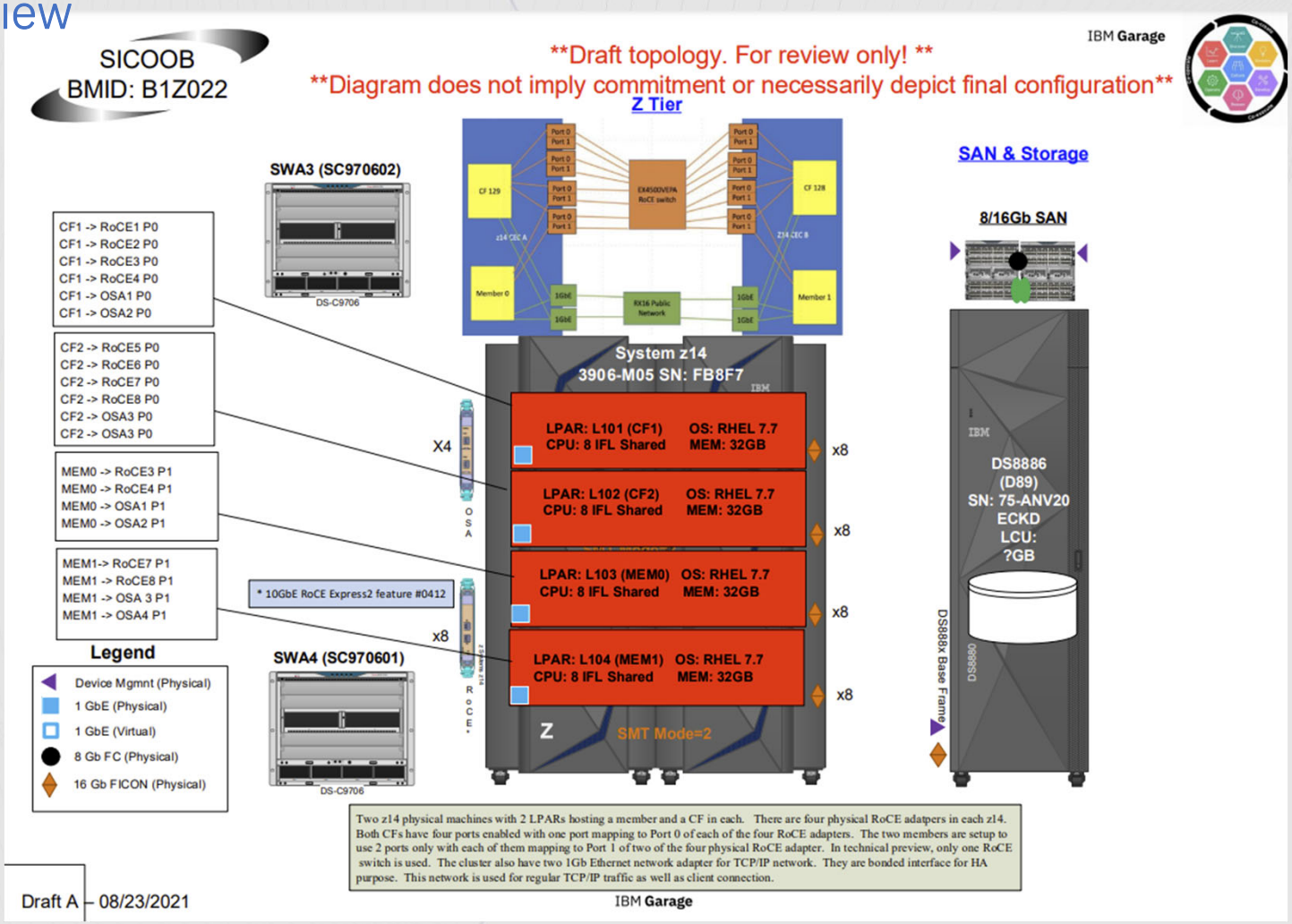
Machine / VM	Hostname	IP
L101 – RHEL 7.7	z207cf1	129.40.16.129
L102 – RHEL 7.7	z207cf2	129.40.16.130
L103 – RHEL 7.7	z207mem1	129.40.16.131
L104 – RHEL 7.7	z207mem2	129.40.16.132
x86 – RHEL 8.4	x207vm1	129.40.16.149

STORAGE:
MTM: 2831-981
Cache: 256GB
Media Type: Enterprise 10K
LCU Assignment: LCU BExx
Volume Assignment: Mod54 BE00 – BE17
Boot devices /var /tmp etc.

IBM Garage - PureScale POT Network

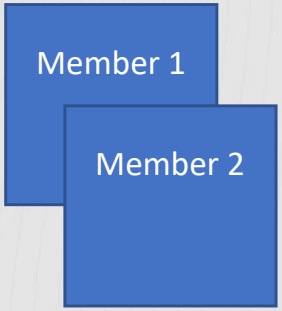


IBM Garage - PureScale POT Overview



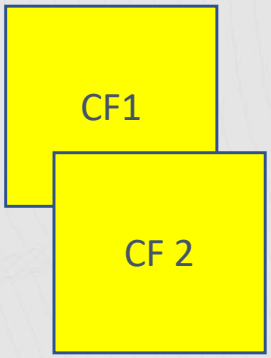
IBM Garage - PureScale POT
Detailed Infrastructure

Linux S390x RHEL 7.7
8 IFLs/16 vCPUs
32GB of Memory



"/"	DASD 3390-9
"/archlogs"	200GB SCSI LUN
"/data"	160GB SCSI LUN
"/primarylogs"	40GB SCSI LUN
db2fs1	40GB SCSI LUN,
tiebreaker disk	1GB

Linux S390x RHEL 7.7
4 IFLs/8 vCPUs
32GB of Memory



Also
2 PVs / 1 VG / 5 LVMs

```
root@db2psf01 dev]# df –h
Filesystem                      Size  Used Avail Use% Mounted on
devtmpfs                         16G     0   16G   0% /dev
tmpfs                            16G  8.0K   16G   1% /dev/shm
tmpfs                            16G   57M   16G   1% /run
tmpfs                            16G     0   16G   0% /sys/fs/cgroup
/dev/dasda1                       7.8G  234M  7.2G   4% /
/dev/mapper/rhel_system-usr          15G  2.1G   13G  15% /usr
/dev/mapper/rhel_system-var         9.9G  4.3G  5.2G  45% /var
/dev/mapper/rhel_system-opt          15G  2.9G   11G  21% /opt
/dev/mapper/rhel_system-home         25G  7.3G   16G  32% /home
/dev/mapper/rhel_system-tmp         8.9G  2.3G  6.3G  27% /tmp
svt3070a.torolab.ibm.com:/vol/alexZ1  450G  442G  8.6G  99% /torolab
tmpfs                            3.2G     0  3.2G   0% /run/user/1005
archlogs                          200G  179G   22G  90% /archlogs
data                             160G   24G  137G  15% /data
db2fs1                            40G  5.0G   36G  13% /db2sd_20210826142048
primarylogs                        40G   24G   17G  60% /primarylogs
tmpfs                            3.2G     0  3.2G   0% /run/user/0
```

```
[root@db2psf01 dev]# pvdisplay
  --- Physical volume ---
PV Name           /dev/dasda2
VG Name           rhel_system
PV Size           <33.27 GiB / not usable <3.11 MiB
Allocatable       yes (but full)
PE Size           4.00 MiB
Total PE          8516
Free PE           0
Allocated PE      8516
PV UUID           2qabg0-xHQs-EsLE-kvae-E3N1-DoQ0-U41HEg
  --- Physical volume ---
PV Name           /dev/dasdb1
VG Name           rhel_system
PV Size           <41.27 GiB / not usable 3.12 MiB
Allocatable       yes
PE Size           4.00 MiB
Total PE          10564
Free PE           261
Allocated PE      10303
PV UUID           gIoruq-IekM-D8vM-VLwZ-XCDL-RYpZ-inS9Zd
```

PoT Networking Details

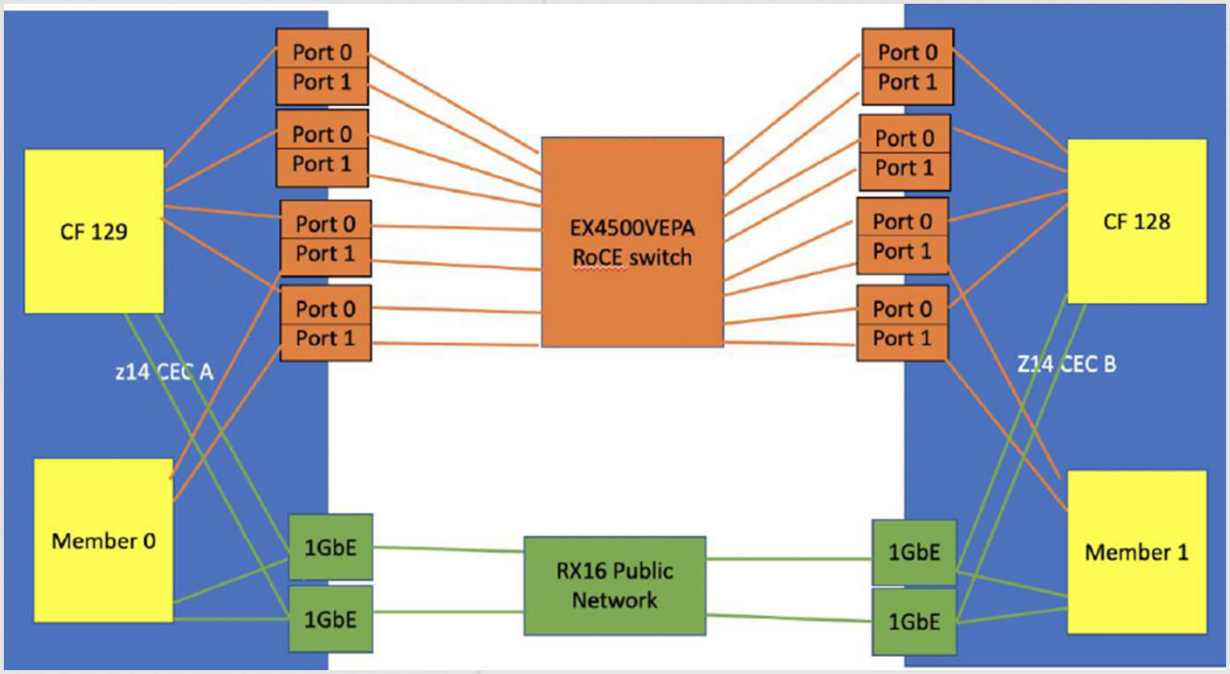
Krvw	IIG	Lqwhuidfh	Lqwhuidfh qdp h	LS Dgguhvv
}53:p hp 4	865	hqS4v865	}53:p hp 40urh3	43141415
	:65	hqv:65	}53:p hp 40urh4	43141515
}53:p hp 5	974	hqS4v974	}53:p hp 50urh3	43141416
	;75	hqv;75	}53:p hp 50urh4	43141516
}53:fi4	445	hqS4v445	]53:fi40urh3	43141417
	644	hqS5v644	]53:fi40urh4	43141517
	645	hqS6v645	]53:fi40urh5	43141617
	:44	hqS7v:44	]53:fi40urh6	43141717
	844	hqS8v844	]53:fi40urh7	43141817
	444	hqv444	]53:fi40urh8	43141917
}53:fi5	755	hqS4v755	]53:fi50urh3	43141418
	554	hqS5v554	]53:fi50urh4	43141518
	555	hqS6v555	]53:fi50urh5	43141618
	;54	hqS7v;54	]53:fi50urh6	43141718
	954	hqS8v954	]53:fi50urh7	43141818
	754	hqv754	]53:fi50urh8	43141918

POT /etc/hosts

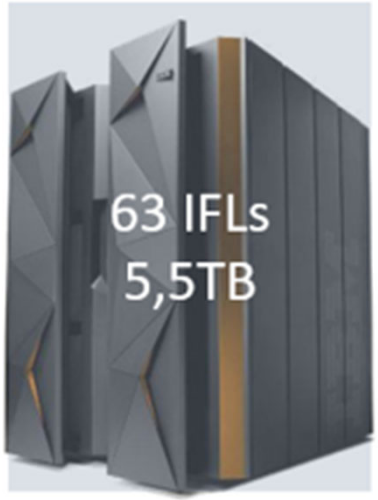
```
[db2sdin1@z207mem1 ~]$ more /etc/hosts

localhost localhost.localdomain localhost4 localhost4.localdomain4
::1         localhost localhost.localdomain localhost6 localhost6.localdomain6

129.40.16.131 z207mem1.pbm.ihost.com  z207mem1
129.40.16.132 z207mem2.pbm.ihost.com  z207mem2
129.40.16.129 z207cf1.pbm.ihost.com  z207cf1
129.40.16.130 z207cf2.pbm.ihost.com  z207cf2

#member 1
10.1.1.2 z207mem1-roe0
10.1.2.2 z207mem1-roe1

#member 2
10.1.1.3 z207mem2-roe0
10.1.2.3 z207mem2-roe1

#cache 1
10.1.1.4 z207cf1-roe0
10.1.2.4 z207cf1-roe1
10.1.3.4 z207cf1-roe2
10.1.4.4 z207cf1-roe3
10.1.5.4 z207cf1-roe4
10.1.6.4 z207cf1-roe5

#cache 2
10.1.1.5 z207cf2-roe0
10.1.2.5 z207cf2-roe1
10.1.3.5 z207cf2-roe2
10.1.4.5 z207cf2-roe3
10.1.5.5 z207cf2-roe4
10.1.6.5 z207cf2-roe5
```

LinuxONE Emperor II –Test Environment for PureScale on Site OI

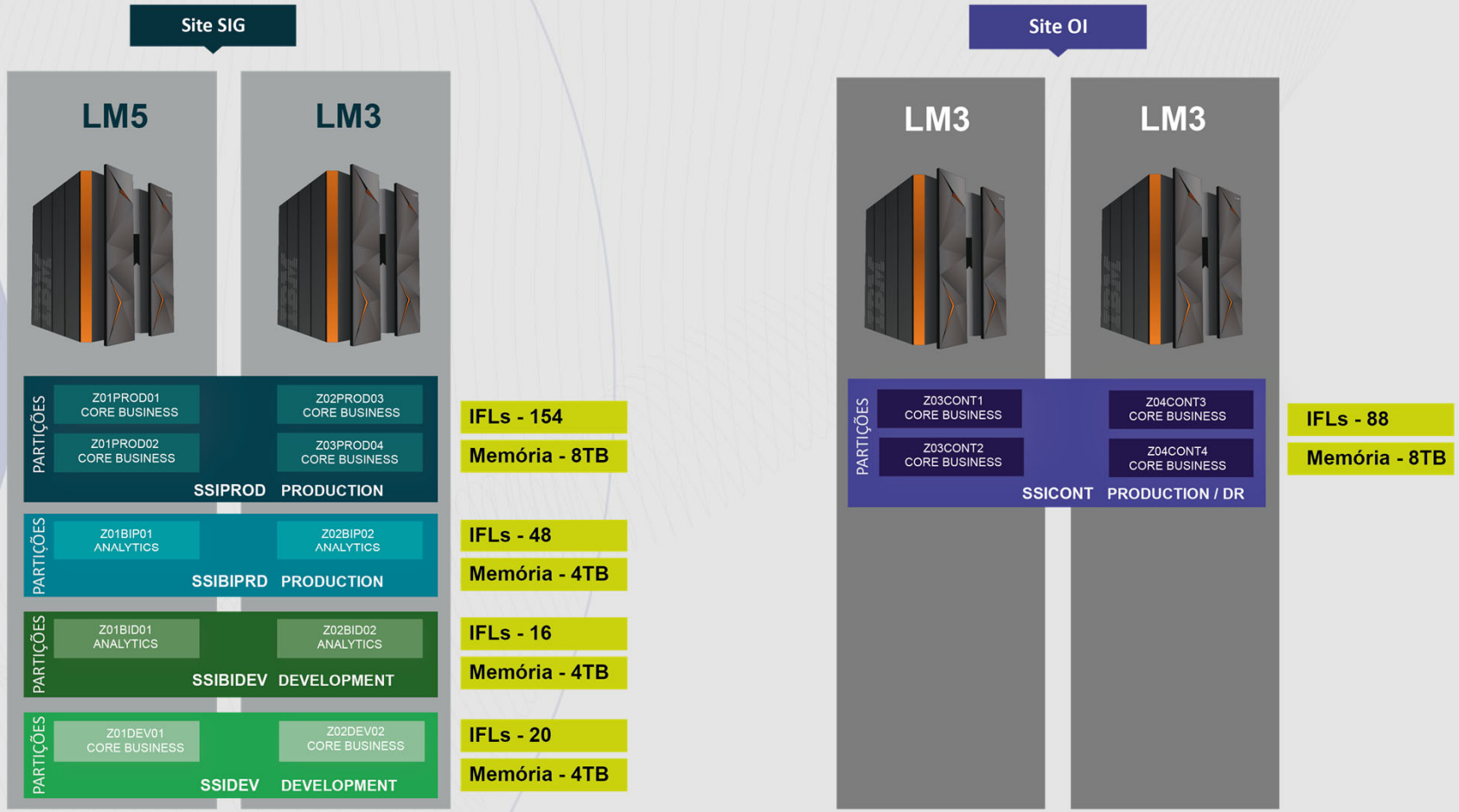
3906-LM3
S/N 822E868
2019



3906-LM3
S/N 820AA28
2018



LinuxONE Emperor II (Z14) – Current Environment



What’s Next

- Upgrade to Z15 by year end
- Setup HADR between the primary and DR site
- Upgrade to Z16 by year end 2024
- SQLServer consolidation
  - 440 SQL Server DBs (1700 X86 cores) into a single pureScale on LinuxONE cluster
  - Each DB contains 2400 objects

Why consolidate onto LinuxONE

- It can replace up to 1,000 x86 cores
- It runs databases 9 times faster than x86
- Runs up to 2.4 million containers
- Manages more than 1 trillion HTTPS transactions (equivalent to half a years' worth of Google searches) in a single day
- It's secure, from both external and internal threats
- Offers an average of 70% reduction in software costs
- 60% reduction in energy/carbon footprint
- Almost 50% overall cost reduction compared with x86
- Designed to never fail with a Mean Time Between Failure of 50+ years

IDUG

Leading the Db2 User Community for 35 Years

IDUG

2023 EMEA Db2 Tech Conference

Db2 pureScale on Z/Linux

Session Code d3(9005)

Please fill out your session evaluation!