NonStop Academy

What is the NonStop Crown Jewel?

A closer look at the NonStop Kernel (NSK)

Bert van Es – Senior Instructor – NonStop Academy

April 2024

Agenda

- Overview of L-series NonStop Operating Systems Architecture – Message based
- How is the x86 Based processor used? Memory Management – Little and Big Endian
- InfiniBand and Cluster I/O Module subsystems
- Debugging and Run-Time architecture – TNS and TNS/X processes - Process control
- Guardian and Open System Services file systems

Part 1

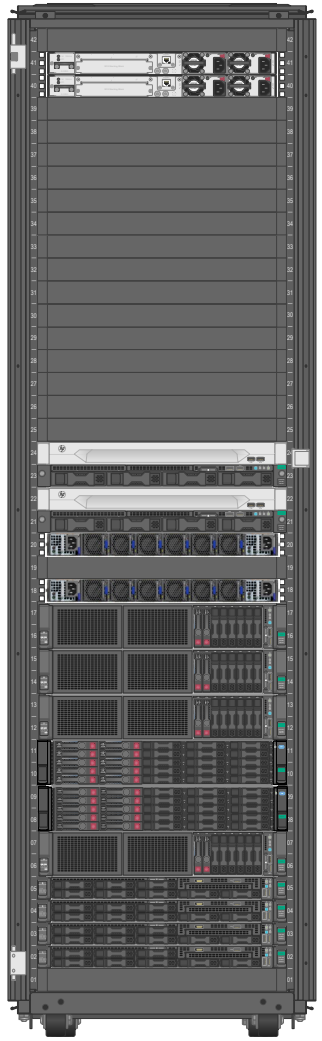
Overview of L-series NonStop Operating Systems Architecture

- Message based

Tandem Computers Design Goals: 1974

- Ability of the node to survive any single hardware or software failure
- Ability of a properly-coded application to survive any single failure
- Assured data integrity
  - When in doubt, fail fast
- Node and application scalability from 2 to 16 processors
- Application object code compatibility across releases
- Online repair of hardware resources
  - Including online reintegration of replacement hardware

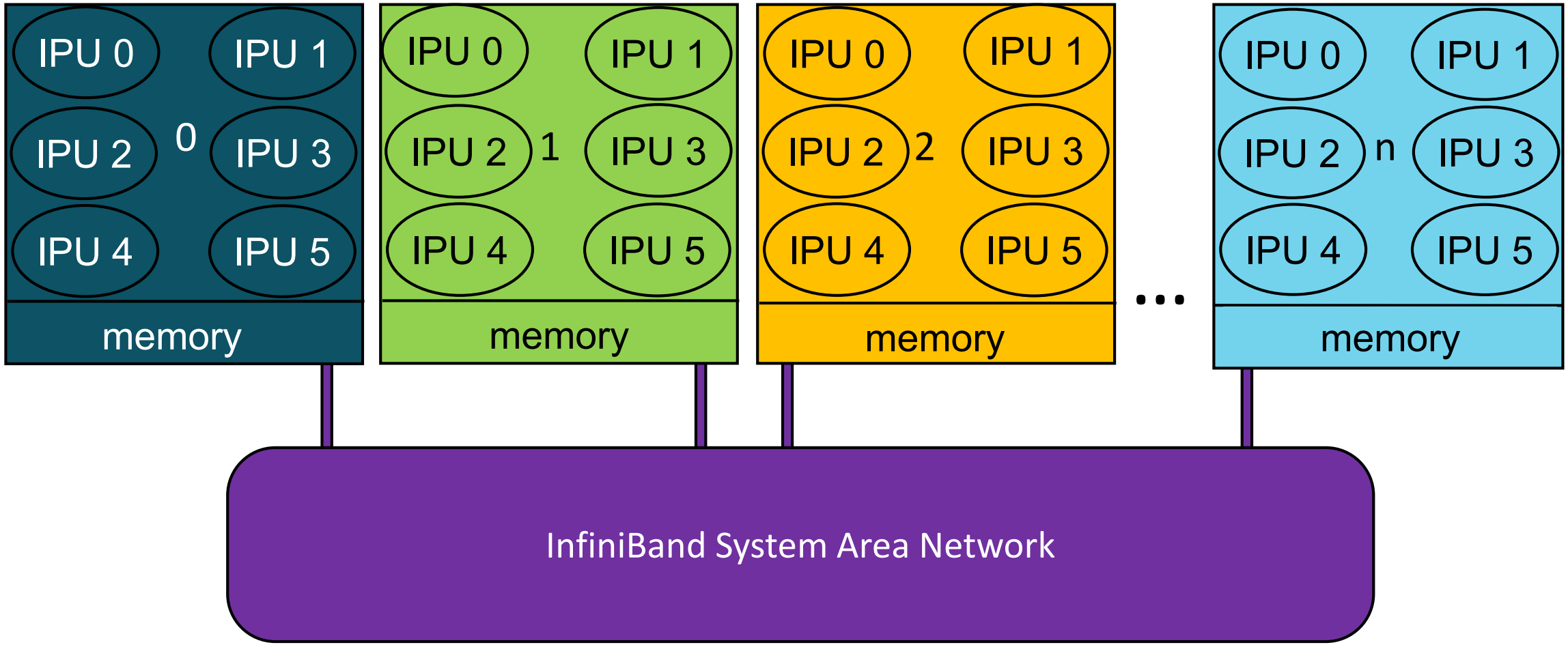
General Characteristics of NonStop Servers



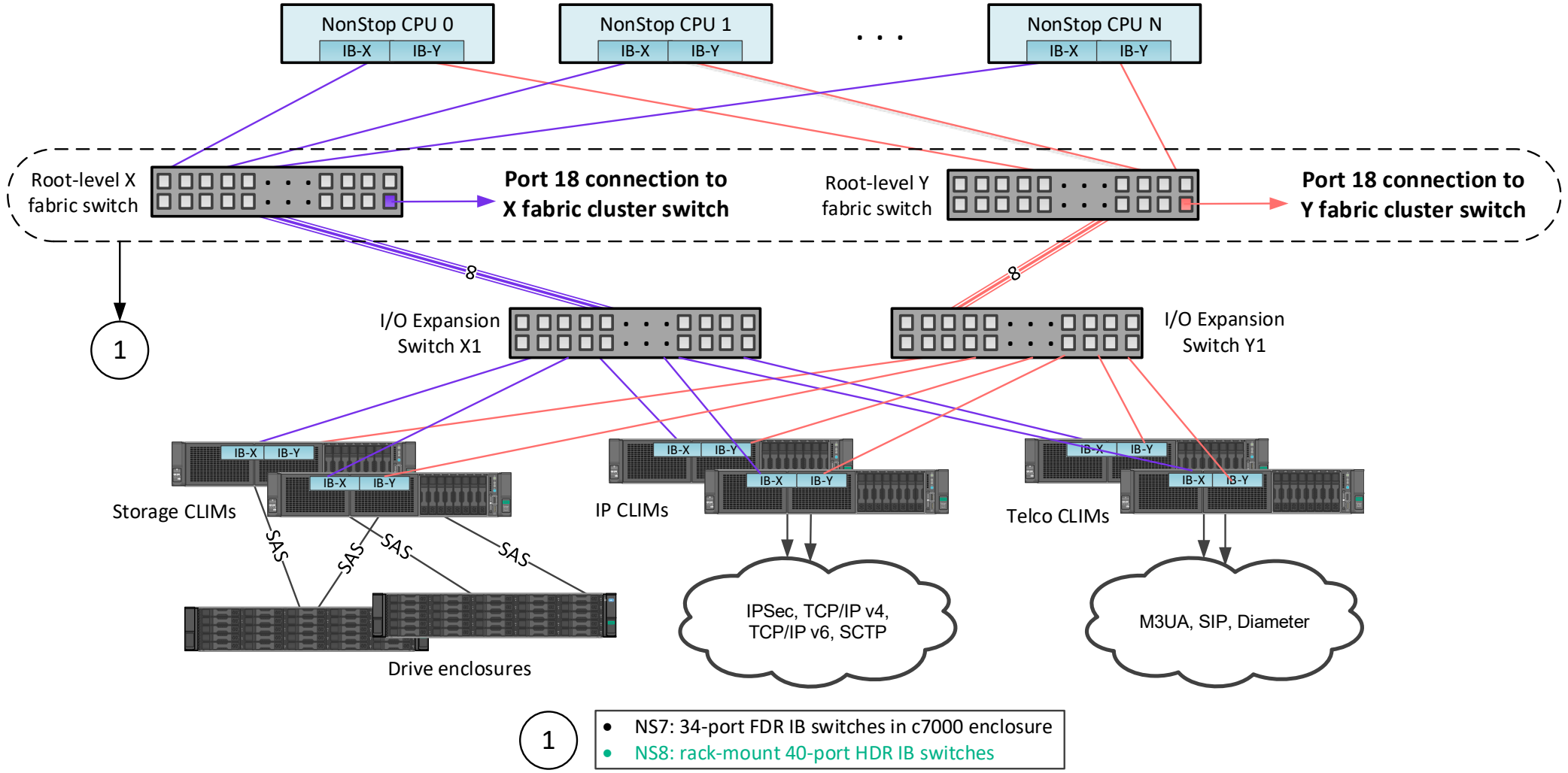
A 4 CPU NS8

- Multiple independent logical processors.
- Hardware fault tolerance.
- Fault-tolerant operating system.
- High-performance SQL database.
- Hardware and software architectures that support:
  - Availability.
  - Data integrity
  - Performance and expandability/scalability
  - Open interface

HPE Integrity NonStop X architecture



NS8 Medium System Interconnect
InfiniBand layout



1

- NS7: 34-port FDR IB switches in c7000 enclosure
- NS8: rack-mount 40-port HDR IB switches

So why is the HPE NonStop OS or NSK the Crown Jewel?

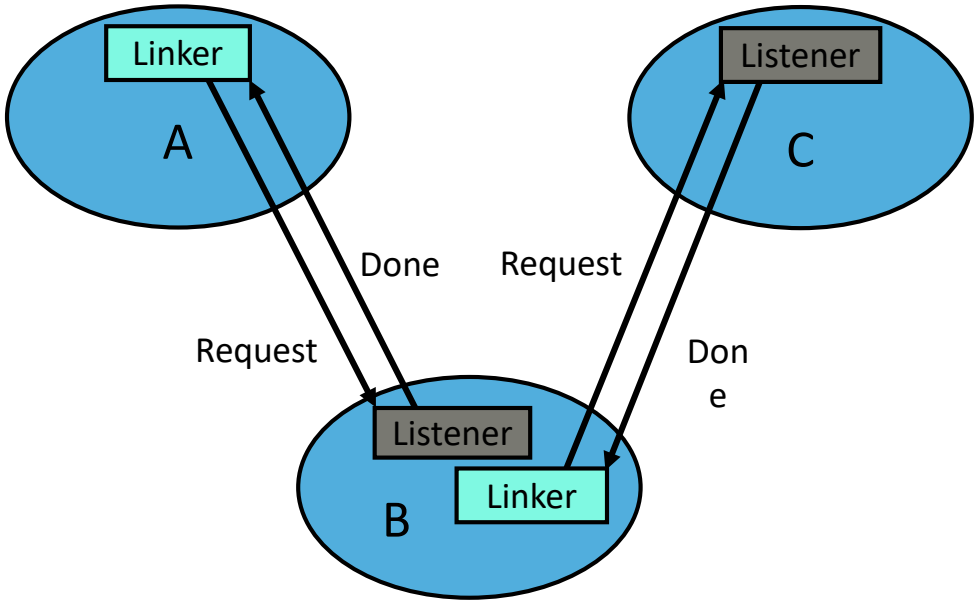
- NSK runs on standard available hardware, so the hardware is not special for NonStop.
- NSK is a message-based operating system giving you:
  - Processor fault tolerance, because each logical processor has its own copy of the operating system not sharing anything with another logical processor, and the message system offers them to communicate with each other and when one fails the other can take over the load, so the application stays up.
  - Software (process) fault tolerance, because the primary process running in one CPU can exchange messages with its backup process with enough information, that when the primary dies, the backup can continue to work where the primary stopped.
  - Scalability – you can start with a small 2-processor system and grow to a 16-processor system and with the same message system over Expand grow to a HPE NonStop network of 4080 processors without to change the application.

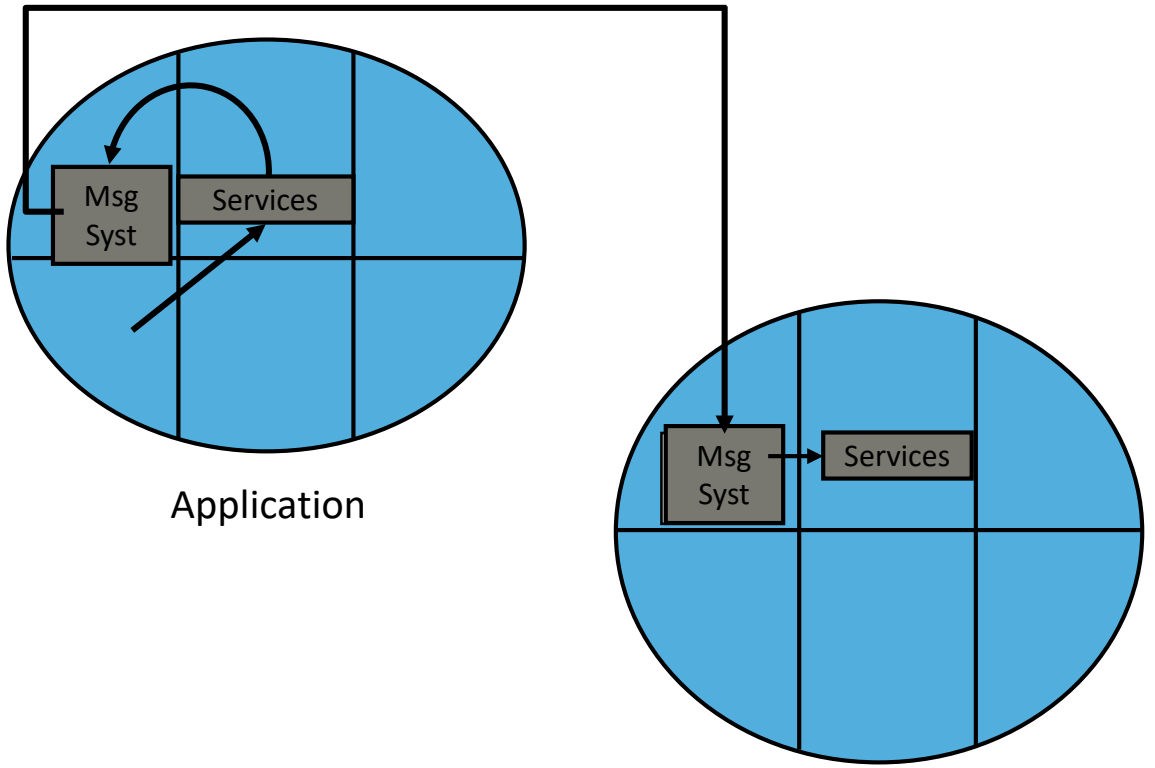
Interprocess Communication

What the Message System Provides

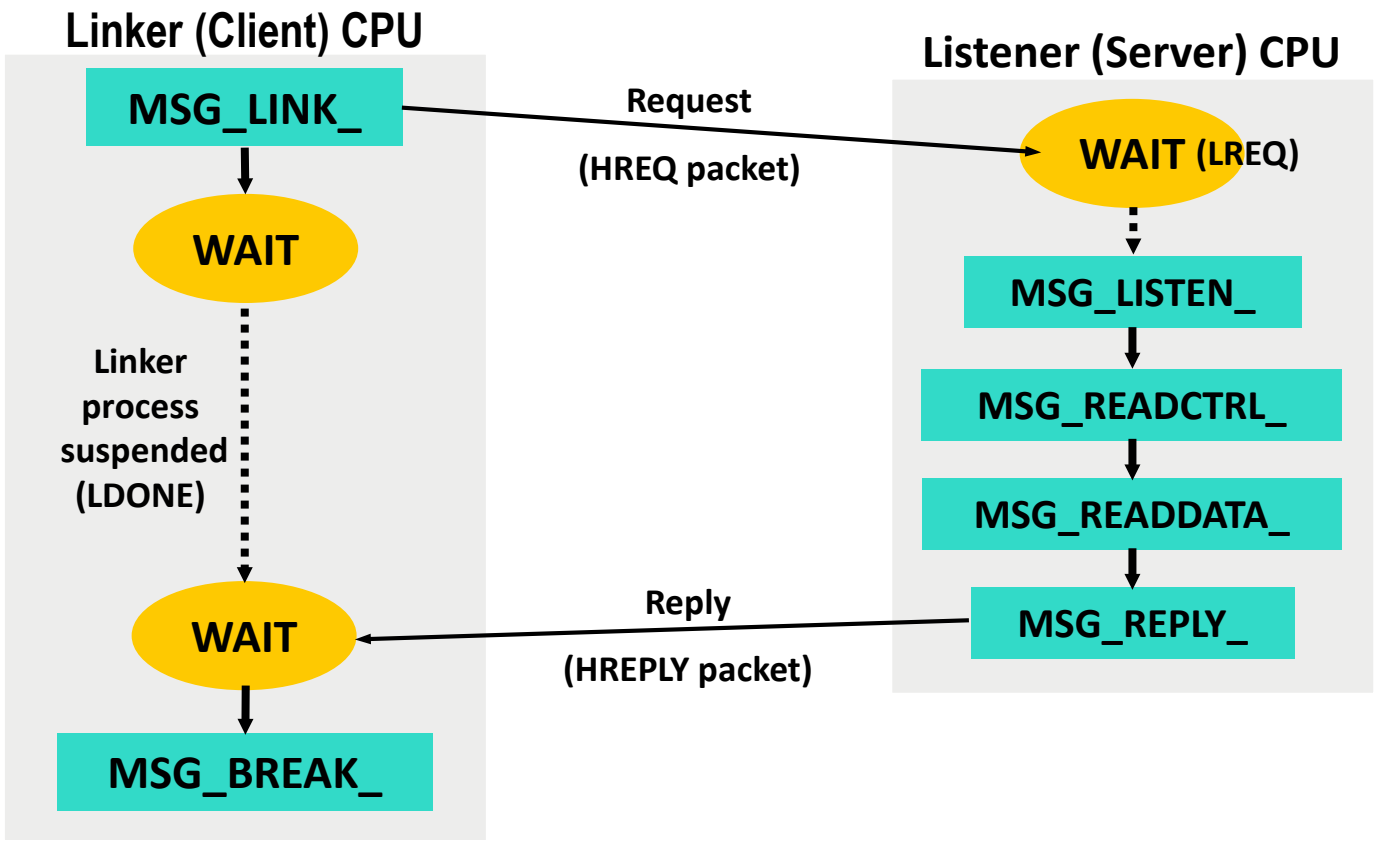
- Geographic independence.
- Fast, priv interface.
- Transport.
- Detects and recovers from errors.
- Message interface is 2-way: Request/reply.
- Nowait interface: Client chooses when to wait for return of incoming message or reply message.
- Sessionless: Unlike the file system, no open or connection establishment necessary. Can always send a message.
- Fault tolerance.

Interprocess Communication — Message System



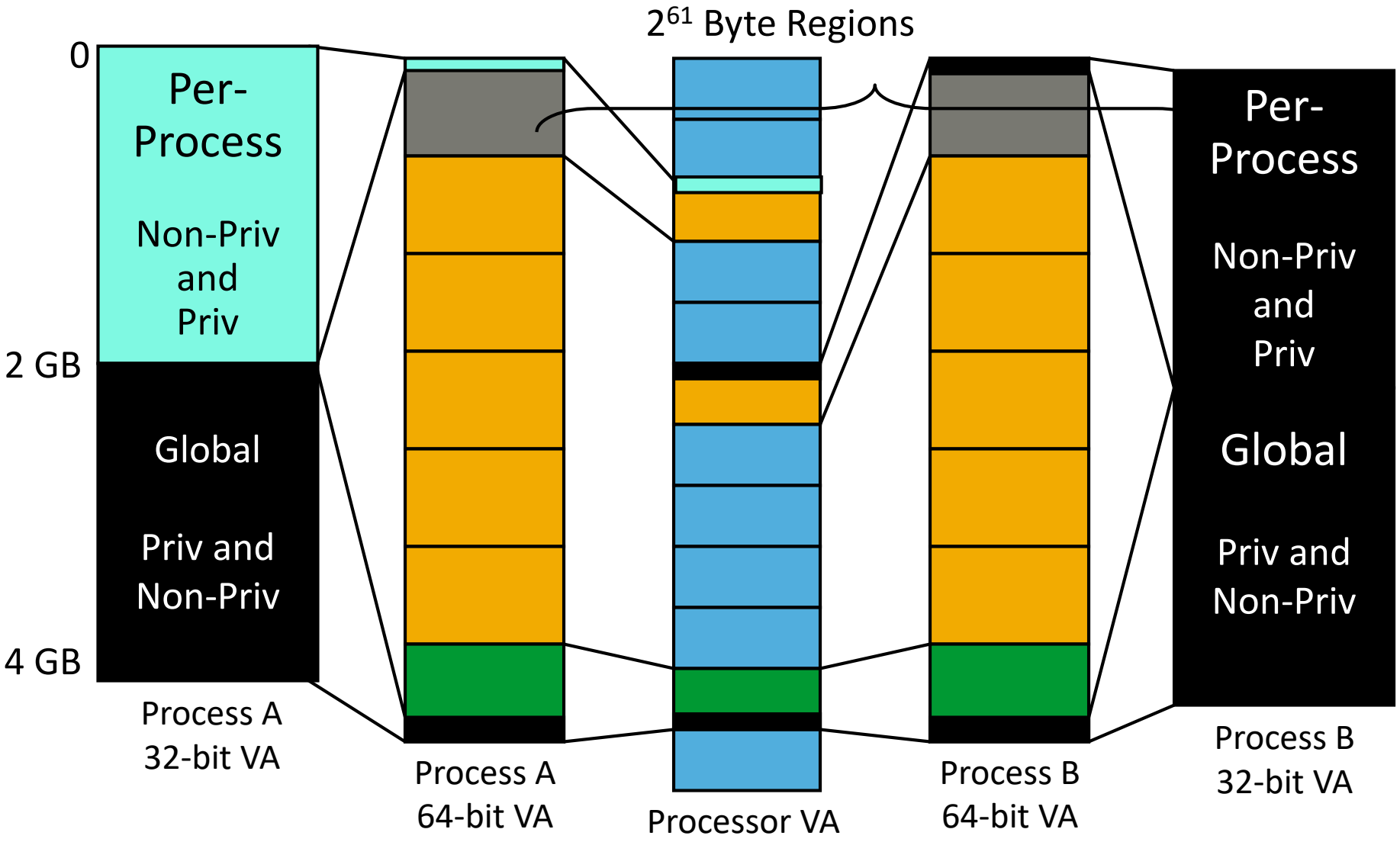


Message System Transfer Protocols (Infiniband)

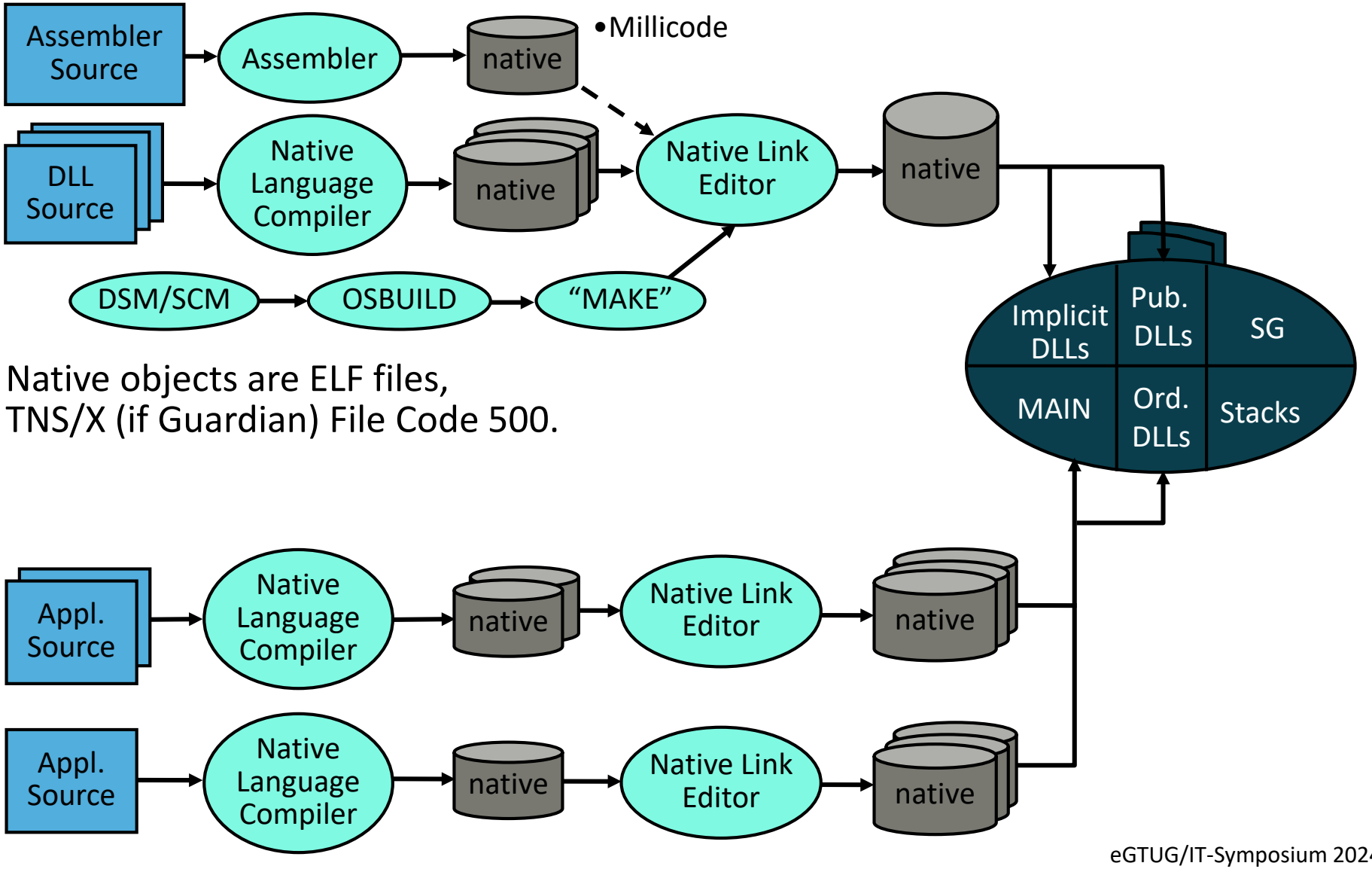


Code Generation, Address Space, Code Space, and Native Process Execution

Virtual Address Spaces

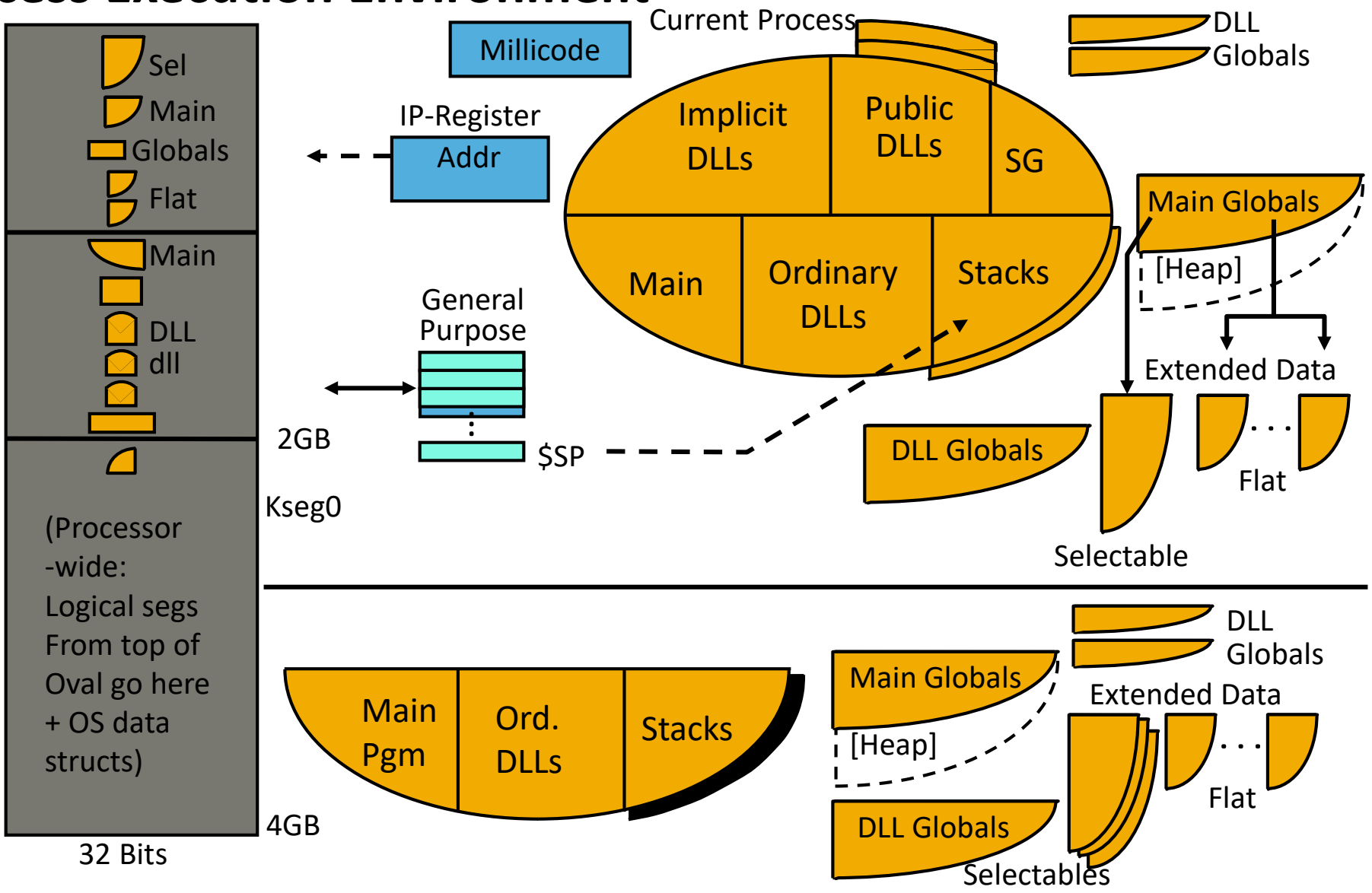


Native Code Generation



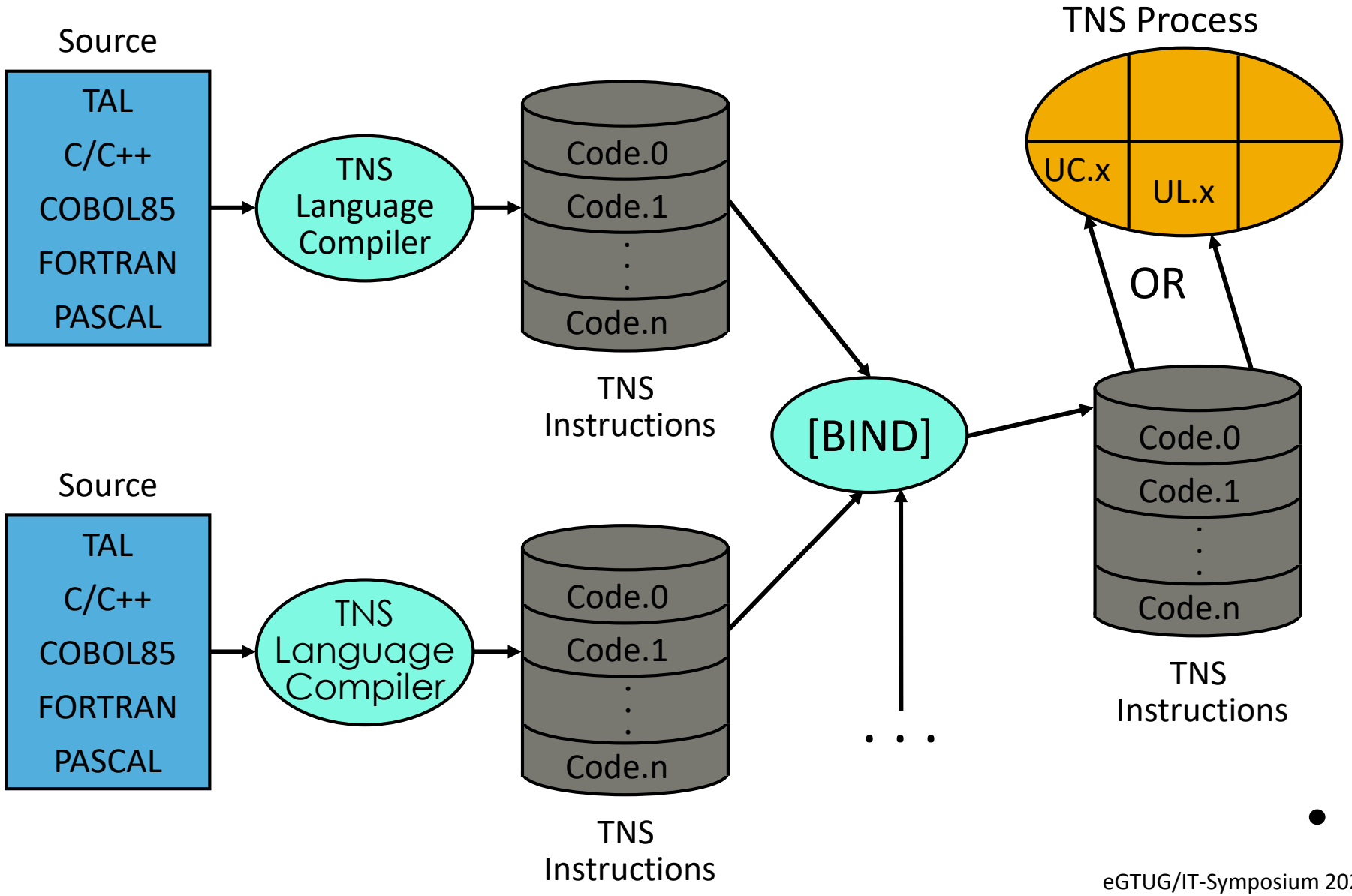
Native objects are ELF files,
TNS/X (if Guardian) File Code 500.

Native Process Execution Environment



Code Generation, Address Space, Code Space, and TNS Process Execution

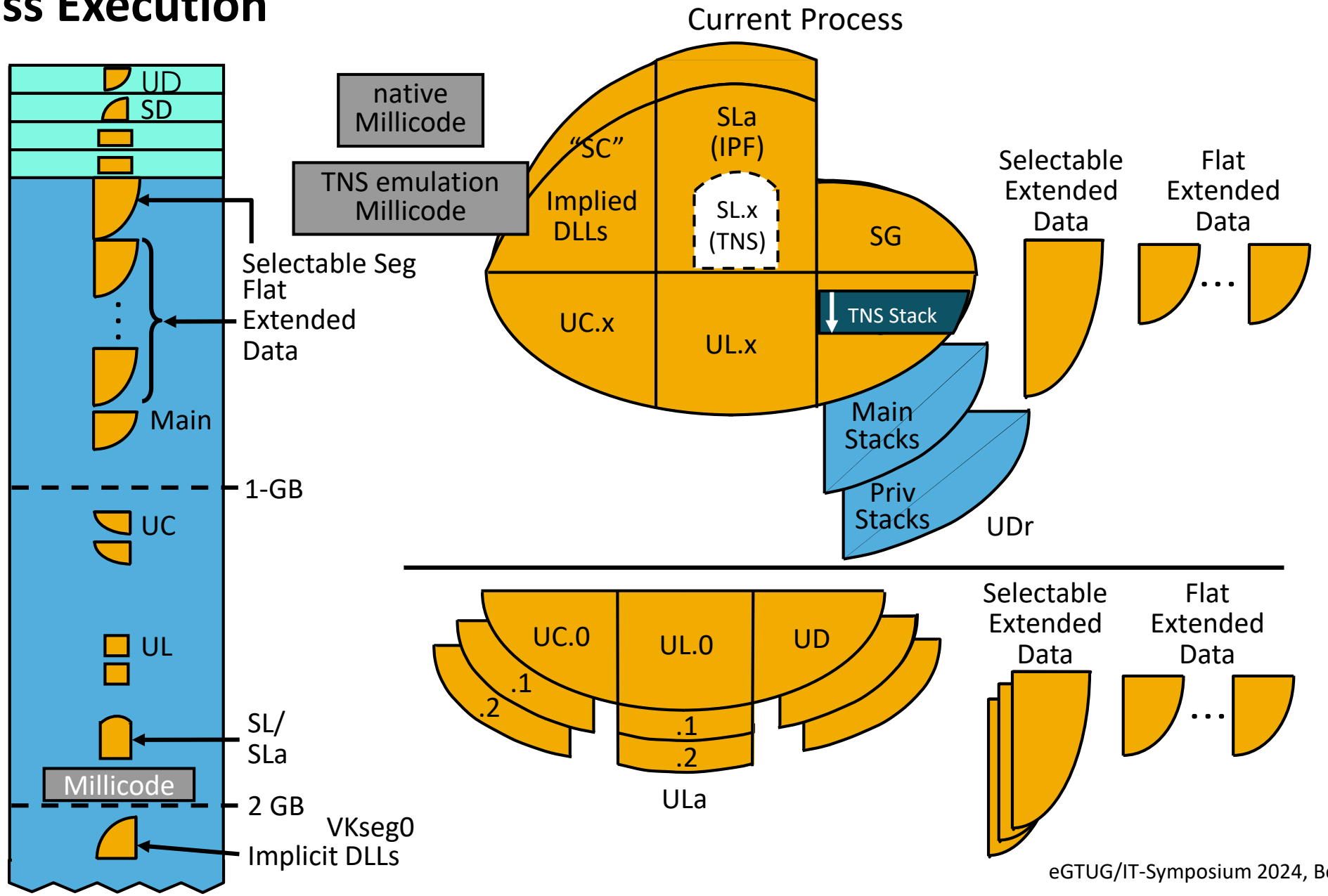
TNS Code Generation



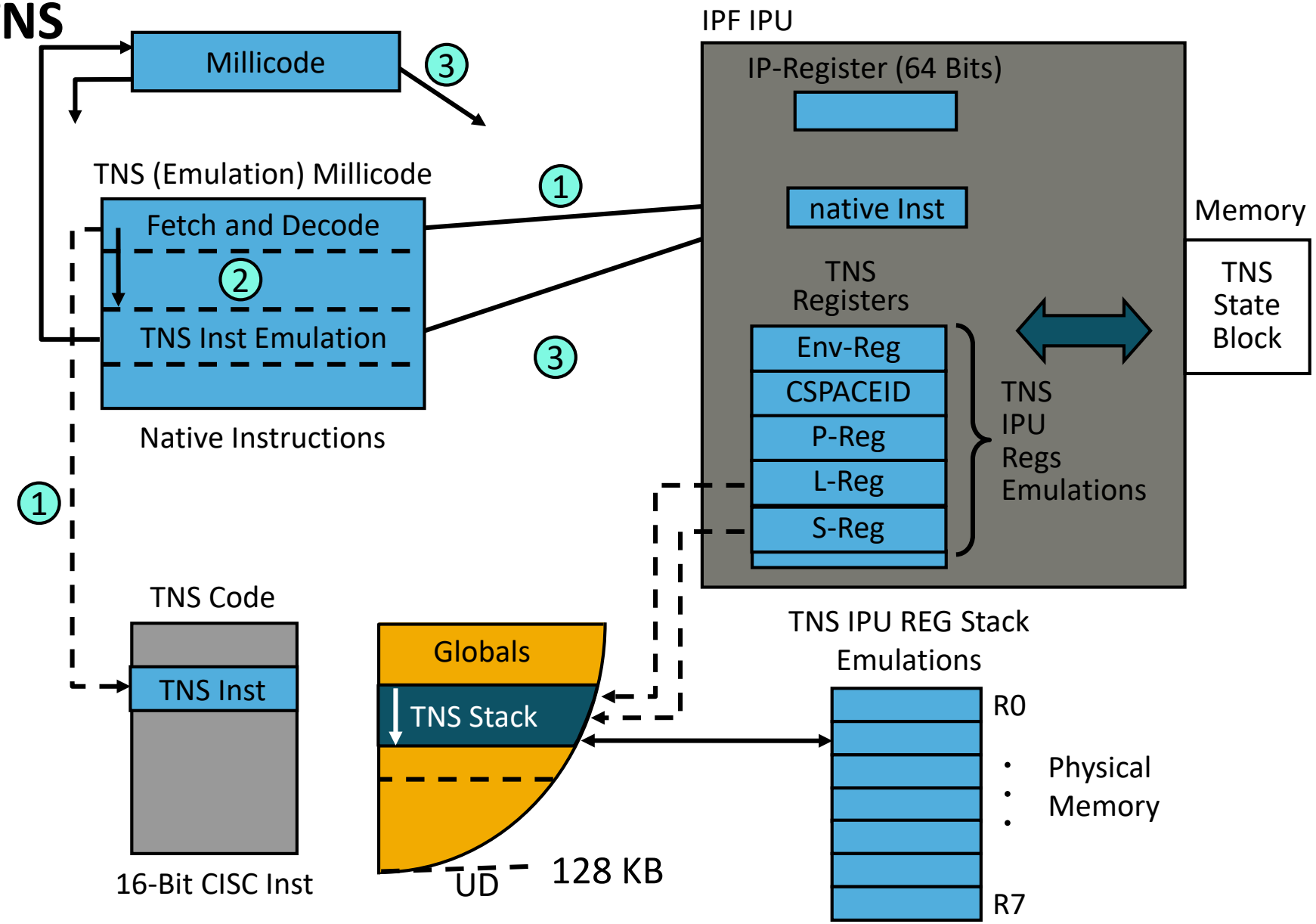
TNS Register Architecture

- ENV register – various status fields
  - 3 bits form the register pointer (which is call RP and frequently used like is was truly separate register)
- Register Stack of 8 16-bit registers, with RP pointing to the current top of stack. This top of stack register is used implicitly by most TNS instructions
- P register – the TNS instruction pointer
- S register – the current top of memory stack register (grows positive)
- L register – the base of the current memory stack frame

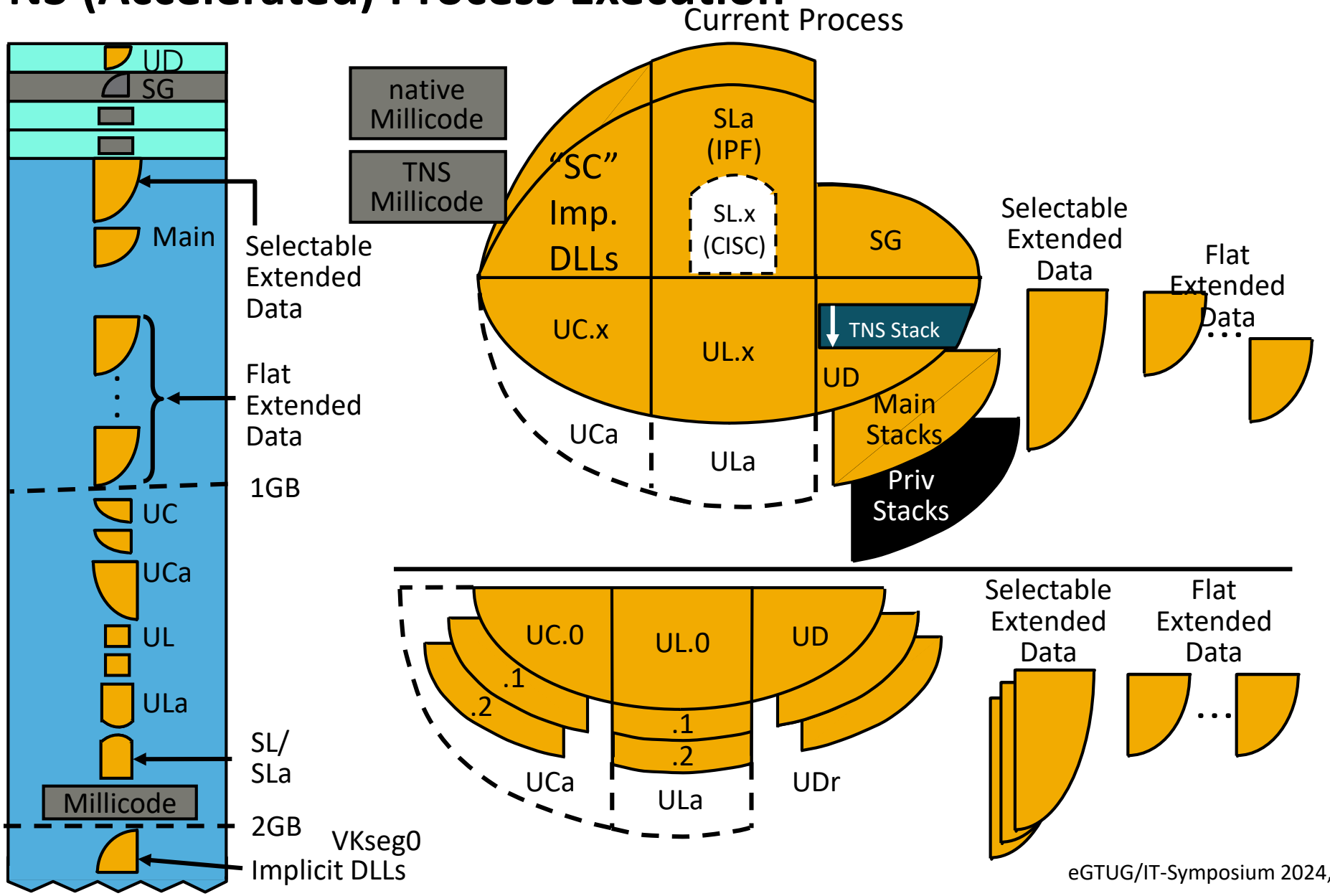
TNS Process Execution



Executing TNS



NonStop TNS (Accelerated) Process Execution



Part 2

How is the x86 Based processor used?

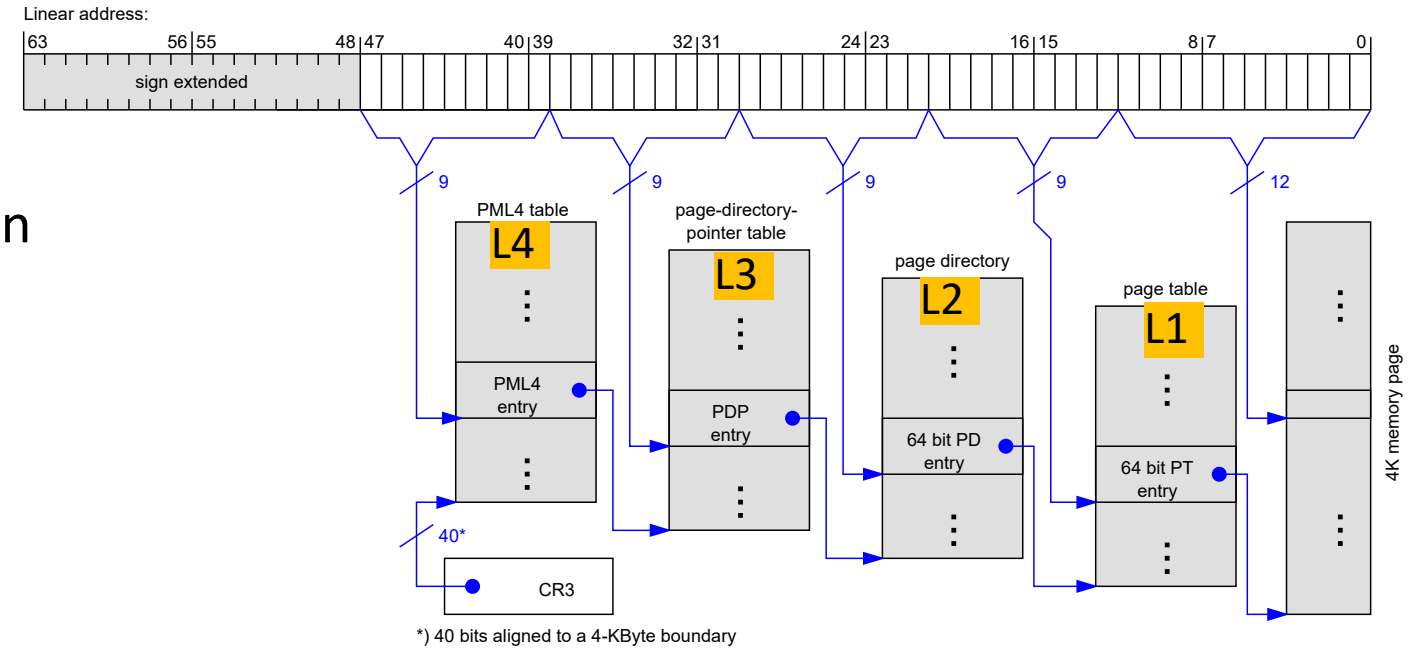
- Memory Management
- Little and Big Endian

Translation Hardware

- x86 is a traditional CISC architecture:
  - The HW does a page table lookup to access the physical page and the TLB caches the result. It may also cache the page tables themselves.
  - Page tables are shared among all IPUs
- The page tables are shared among all IPUs of the CPU.
  - So, XPF uses the GS (or %gs) register for that purpose.
- The system always runs in 64-bit mode and uses four level page tables.

X86-64 Page tables, 4KB pages

- A canonical address is x86-64 is where the 52-48 bit virtual address is sign extended.
- To make a 16KB page 4 4KB PTEs are updated as a group (interrupts masked). The dirty and other bits when checked are the orring of the four entries.
- The TLB is loaded by the HW when a lookup is done.
- Each of the four tables (L4, L3, L2, and L1) have 512 entries.



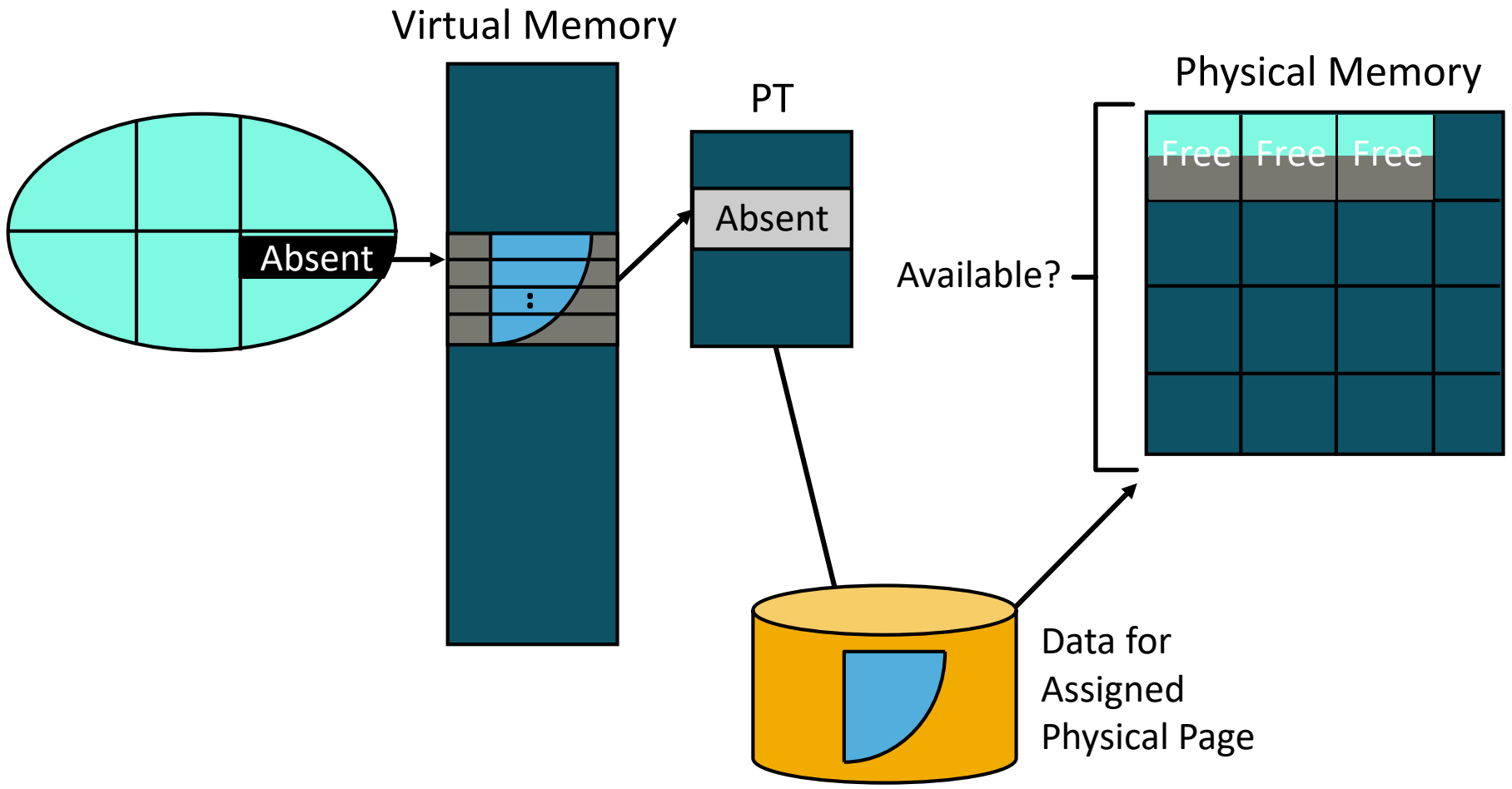
Page Faults

Lesson 8

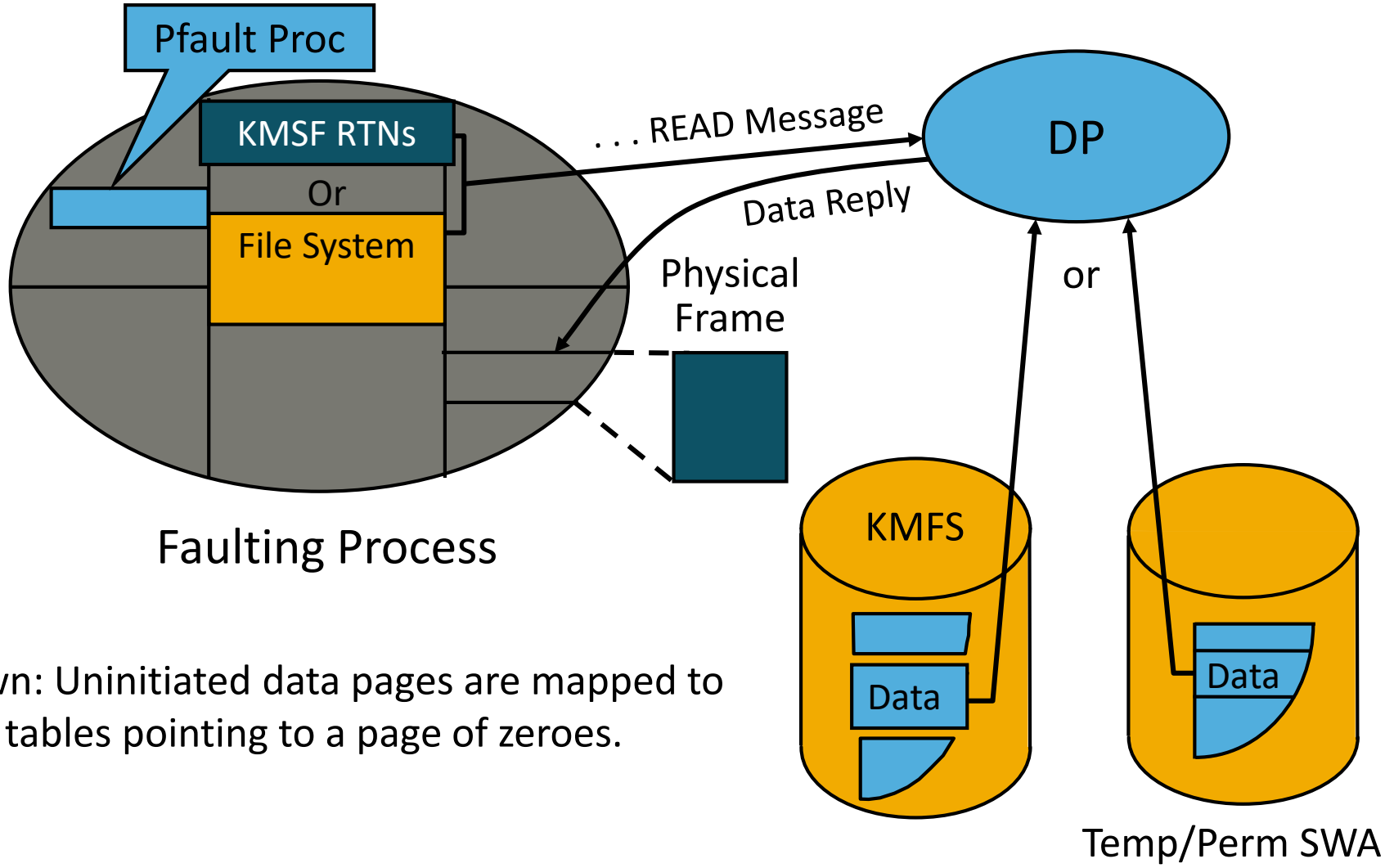
XPF: Software Interrupt Vector Code

- A TLB miss causes a HW page table search which finds an invalid page (for this current kind of access).

Absent — Page Fault

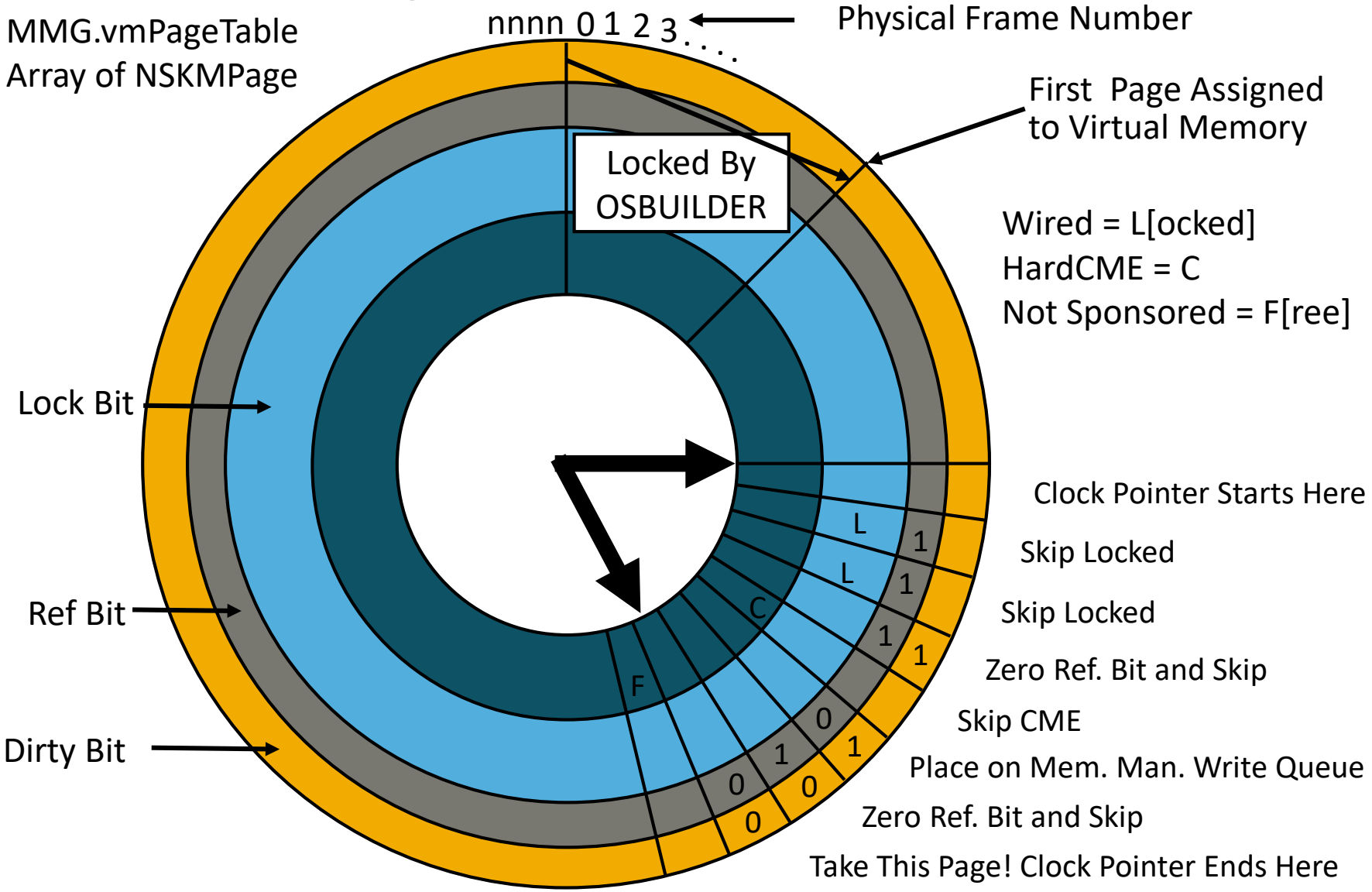


Pfault Proc Initializes the Fault-In Page in Faulting Process Context



Not shown: Uninitiated data pages are mapped to mapping tables pointing to a page of zeroes.

Frame Selection — Clock Algorithm



Memory Manager Process Role Reduced

- Page faults are handled in the context of the faulting process.
- The functions of the memory manager process have been drastically reduced. The memory manager now:
  - During initialization, common opens implicit DLLs.
  - Replenishes the supply of pages available for allocation under mutex.
  - Cleans pages.
- Every 10 seconds, runs the clock to:
  - Age pages (turn off reference bits).
  - Add dirty pages to the dirty page queue.
- Woken up on INTR when there are dirty pages.
  - Writes them out to disk and puts them on a “stealable” list.
  - Can wake up a process waiting for a page.

Endianism - Big vs Little endian

Big vs Little Endian

- Big Endian means that for a given datum the high order byte is stored first in memory (lowest address) and the low order byte is stored last (highest address).
- Little Endian means that for a given datum the low order byte is stored first in memory (lowest address) and the high order byte is stored last (highest address).
- So, for the number x01020304:

address	0000	0001	0002	0003
Big	01	02	03	04
Little	04	03	02	01

- TNS, MIPS, and Itanium are Big endian.
- X86-64 is Little endian

Making XPF Big Endian

- NonStop runs x86 in “Big Endian by doing the following:
  - All program (source level) variables are written to memory as big endian datums.
  - For data that is little endian in memory but must be presented to a user/program the data is changed to big endian format.
- Some parts of the system actually run in little-endian mode (i.e., it sees real little-endian data). This include the TNS emulator and millicode.
- So, if you look at raw memory of a process there is a mix of little- and big-endian items.
- Some things in little endian format:
  - Page table data.
  - Addresses pushed onto the memory stack by a call or interrupt.
  - Register values in a jump buffer.

What parts of the system know about Little endian?

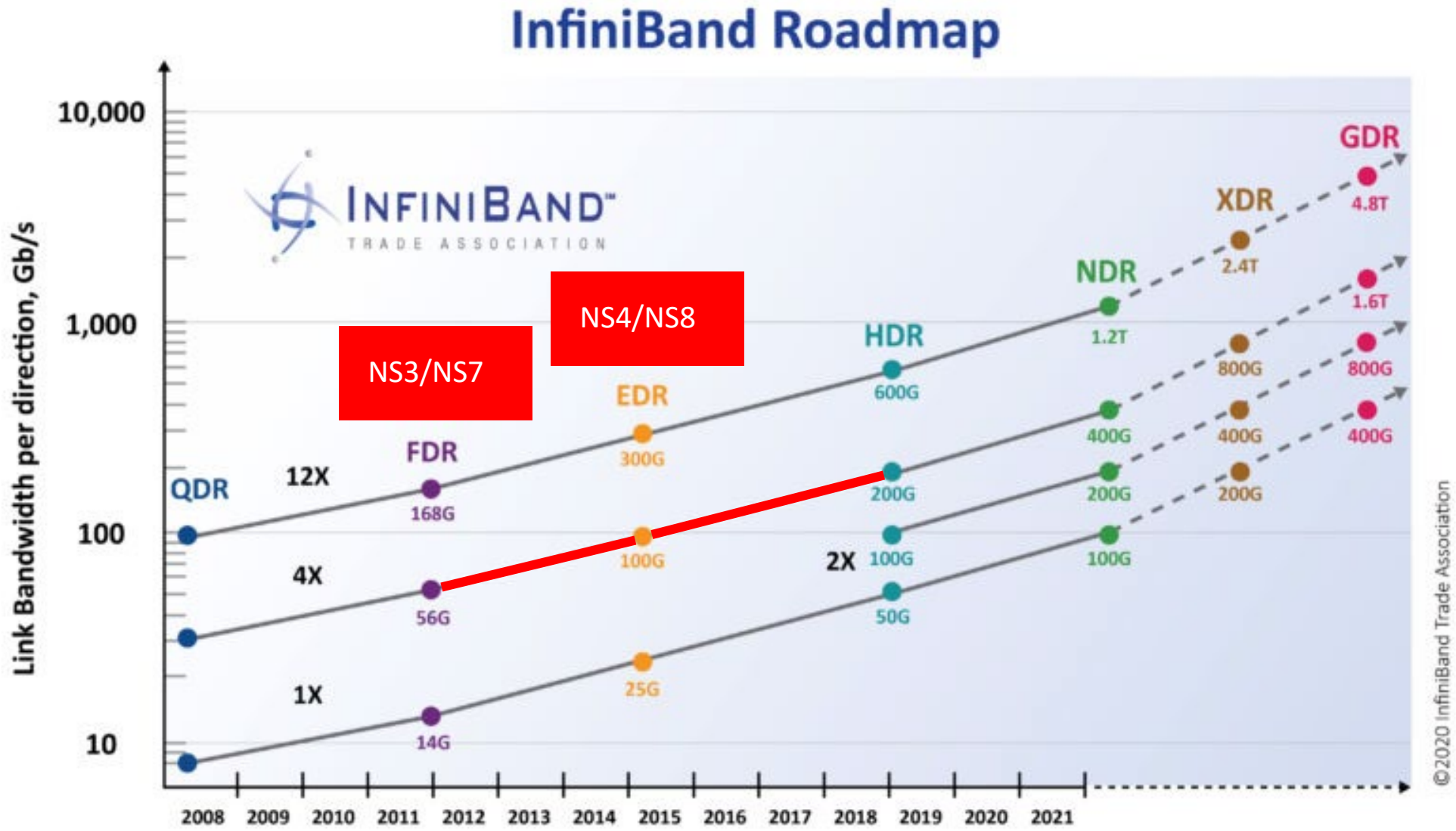
- The OS part that deals with page tables, and other structures known to the x86-hardware.
- The debuggers which know the difference between program variables and “hidden” data.
- The TNS emulator.
- All native compilers.

Part 3

- InfiniBand
- Cluster I/O Module subsystems

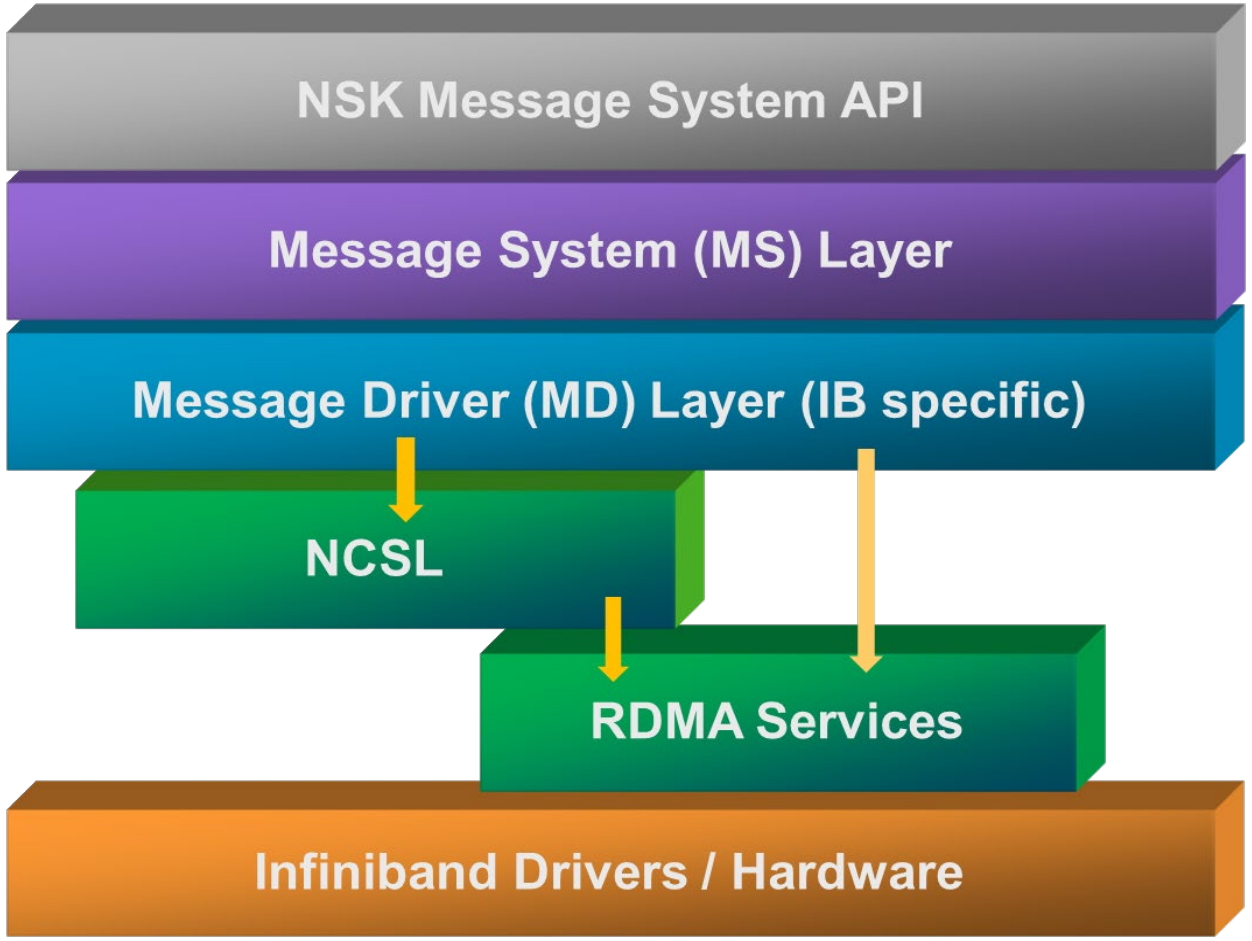
InfiniBand

Fourteen Data Rate InfiniBand (4X)



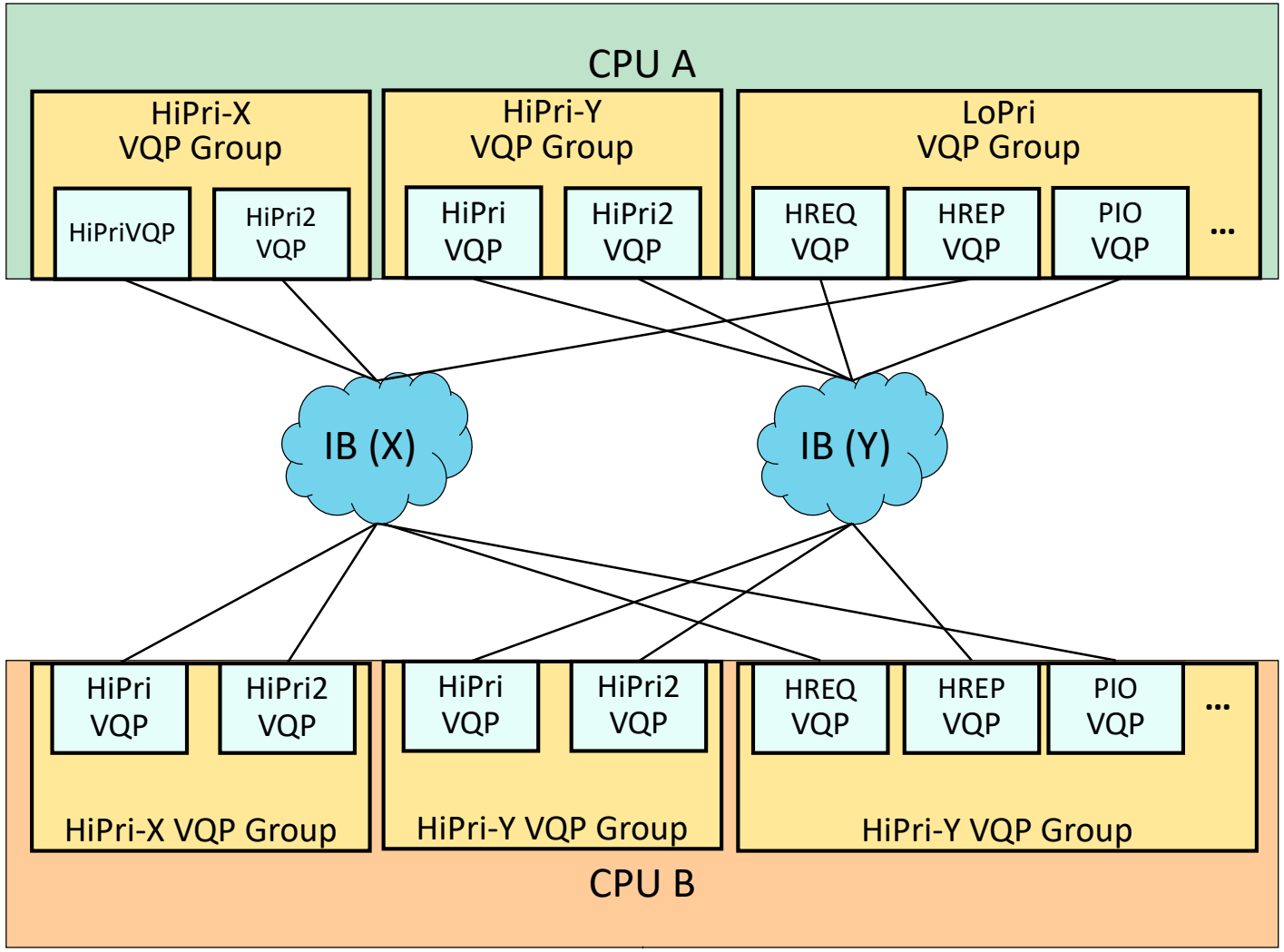
XPF Message System layers

On InfiniBand the Message System has a similar set of layers, with the Message Driver layer changed to support IB and talk to the NCSL API and RDMA Services below it instead of the TNet Services:



IB Connections vs SNet Connnectionless

- Using SvNet each message includes the address of the target CPU.
- However, IB is a connected protocol, so each CPU is connected to every other CPU. (A la sockets)
  - Moreover, there are multiple connections to allow the different classes of messages to move independently.

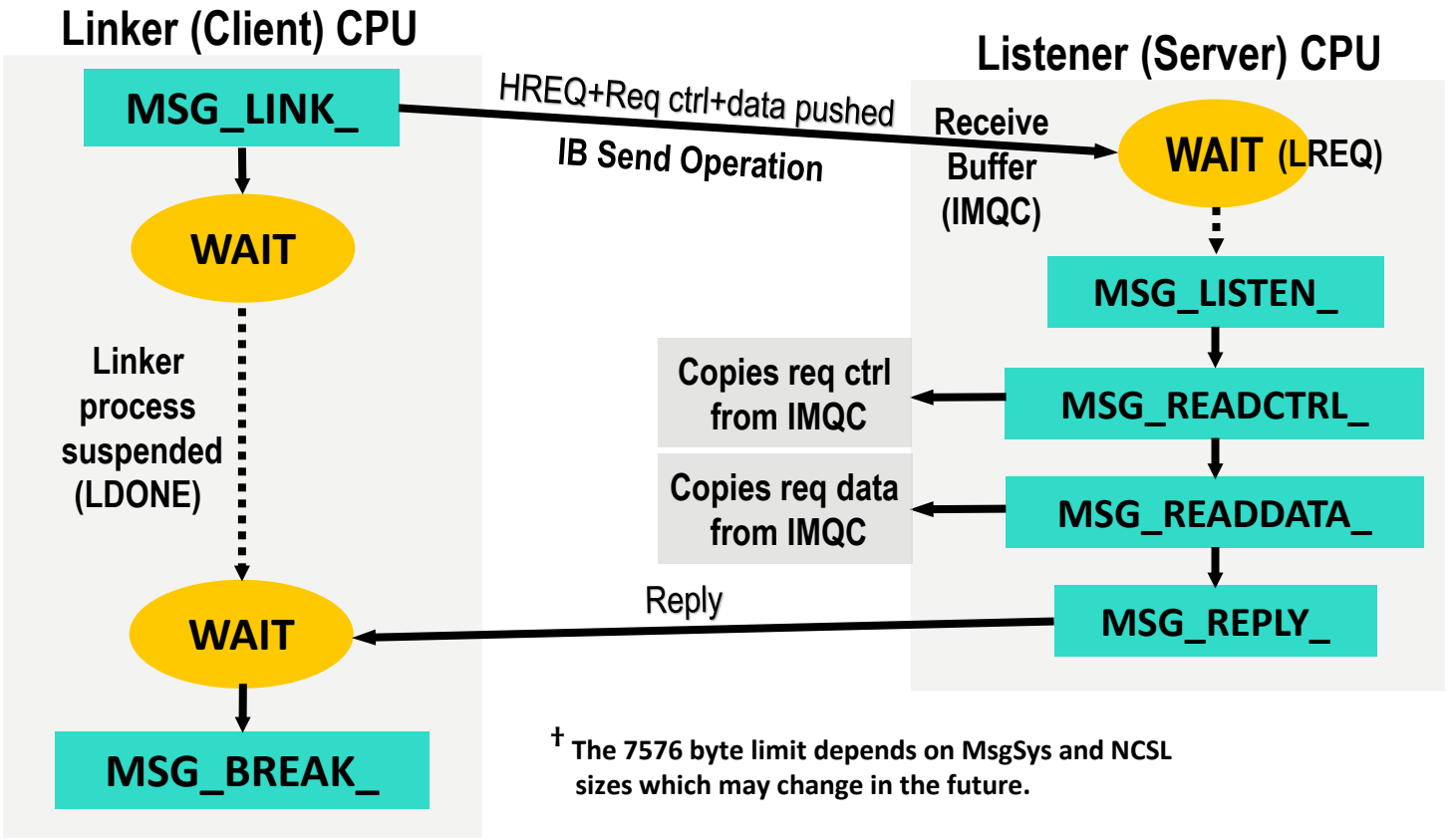


Message System Transfer Protocols on IB

Sending request : ctrl + data size ≤ 7576† bytes

(All the data is pushed along with the HREQ packet in an IB Send operation)

- The generic MS link operation is mapped to an IB send
- The MS reply is also linked to and IB send.



† The 7576 byte limit depends on MsgSys and NCSL sizes which may change in the future.

General NonStop I/O

- I/O transfer:
  - I/O as remote memory transfer.
  - NonStop L-series: InfiniBand uses IB connections and its memory mappings.

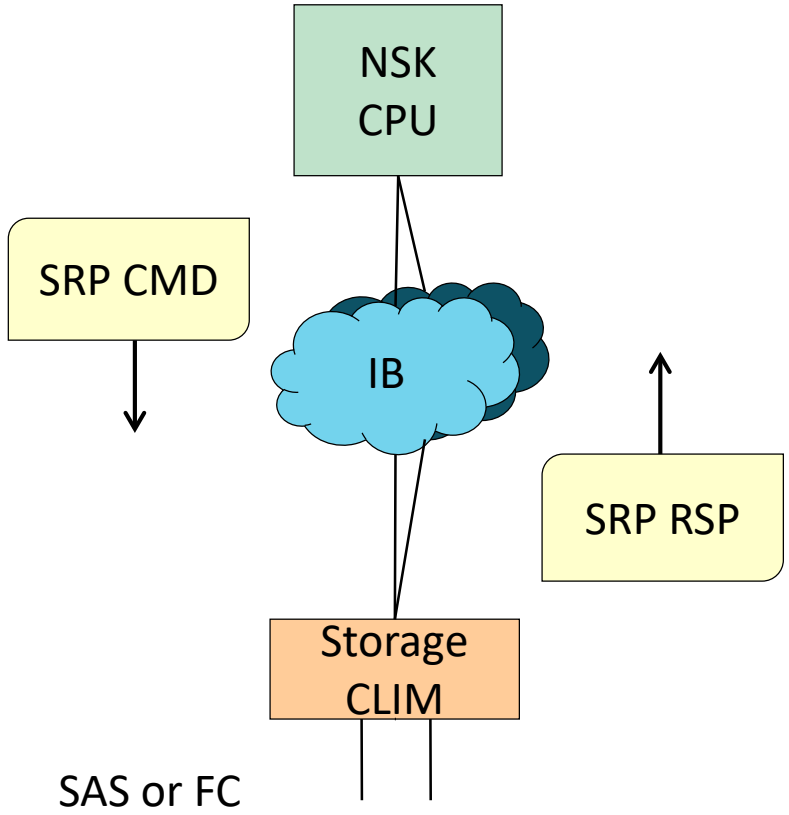
Storage I/O

- High level view of storage I/O
- Life of an I/O operation
  - Initiation
  - Data transfer
  - Completion
- Starting a device

IB based Storage Subsystem

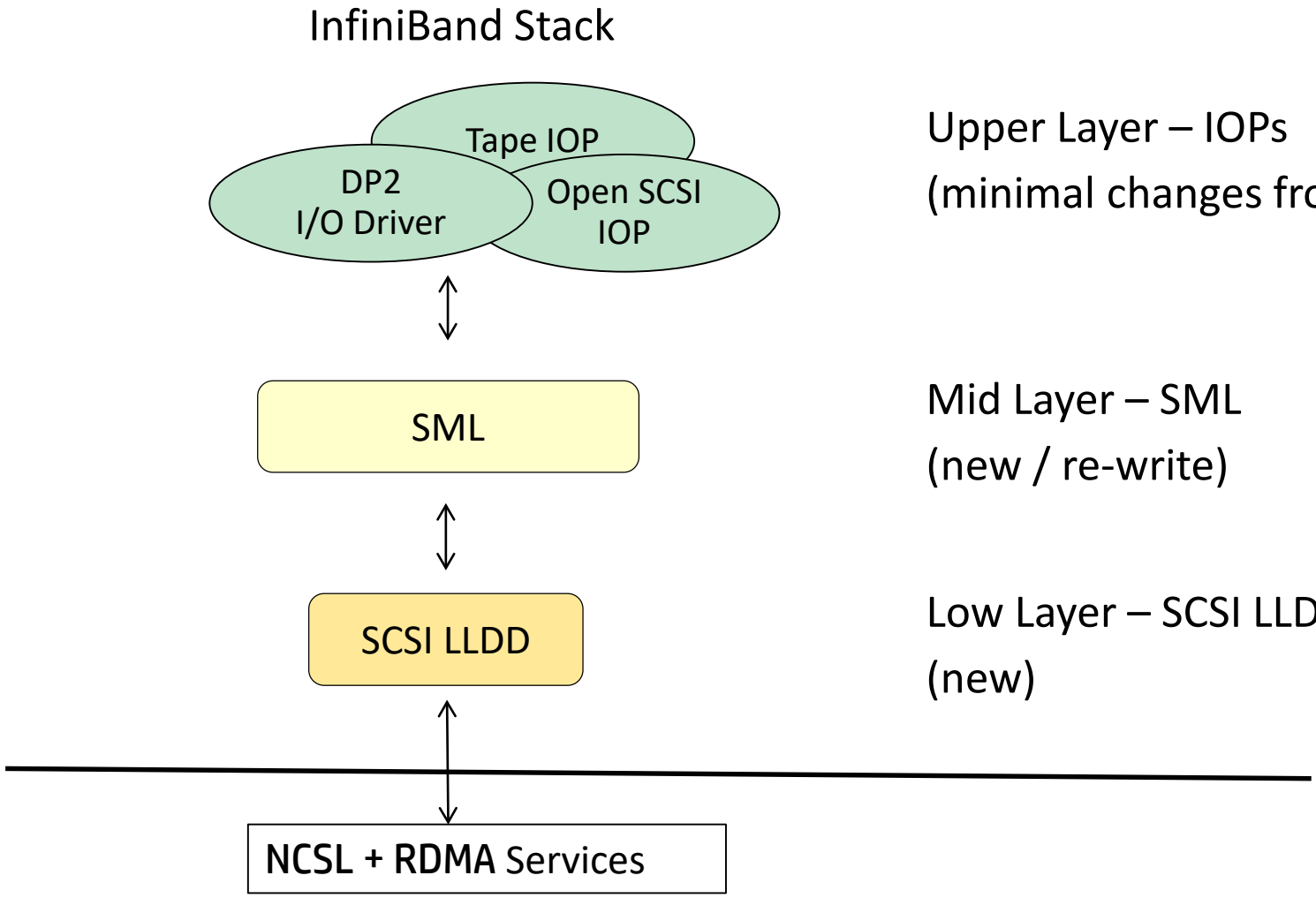
Storage Protocol

InfiniBand: SCSI RDMA Protocol (SRP)



SRP = SCSI RDMA Protocol

XPF Storage Software Stack

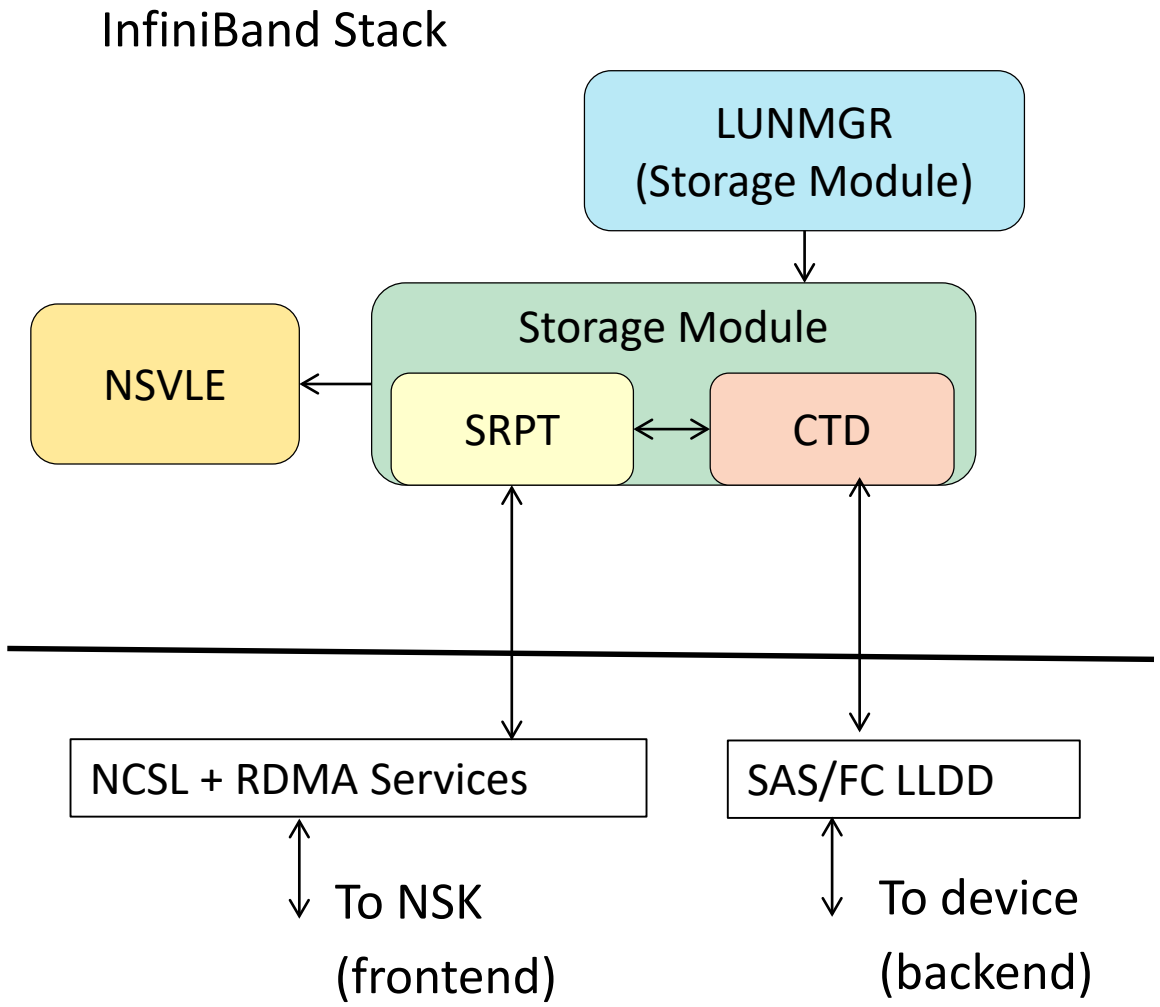


Upper Layer – IOPs
(minimal changes from SNet)

Mid Layer – SML
(new / re-write)

Low Layer – SCSI LLDD
(new)

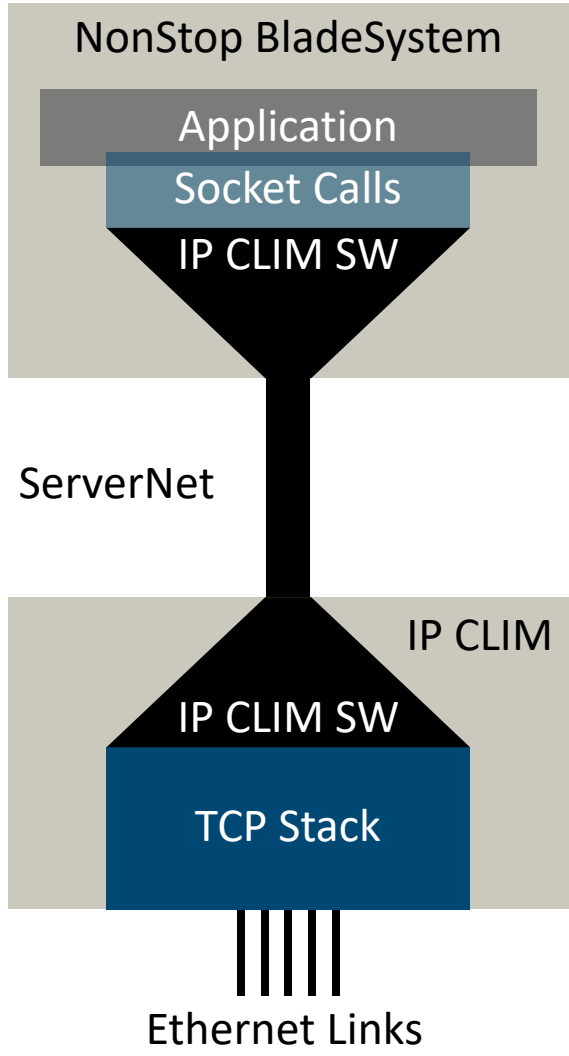
Storage CLIM - CLIM Software Stack



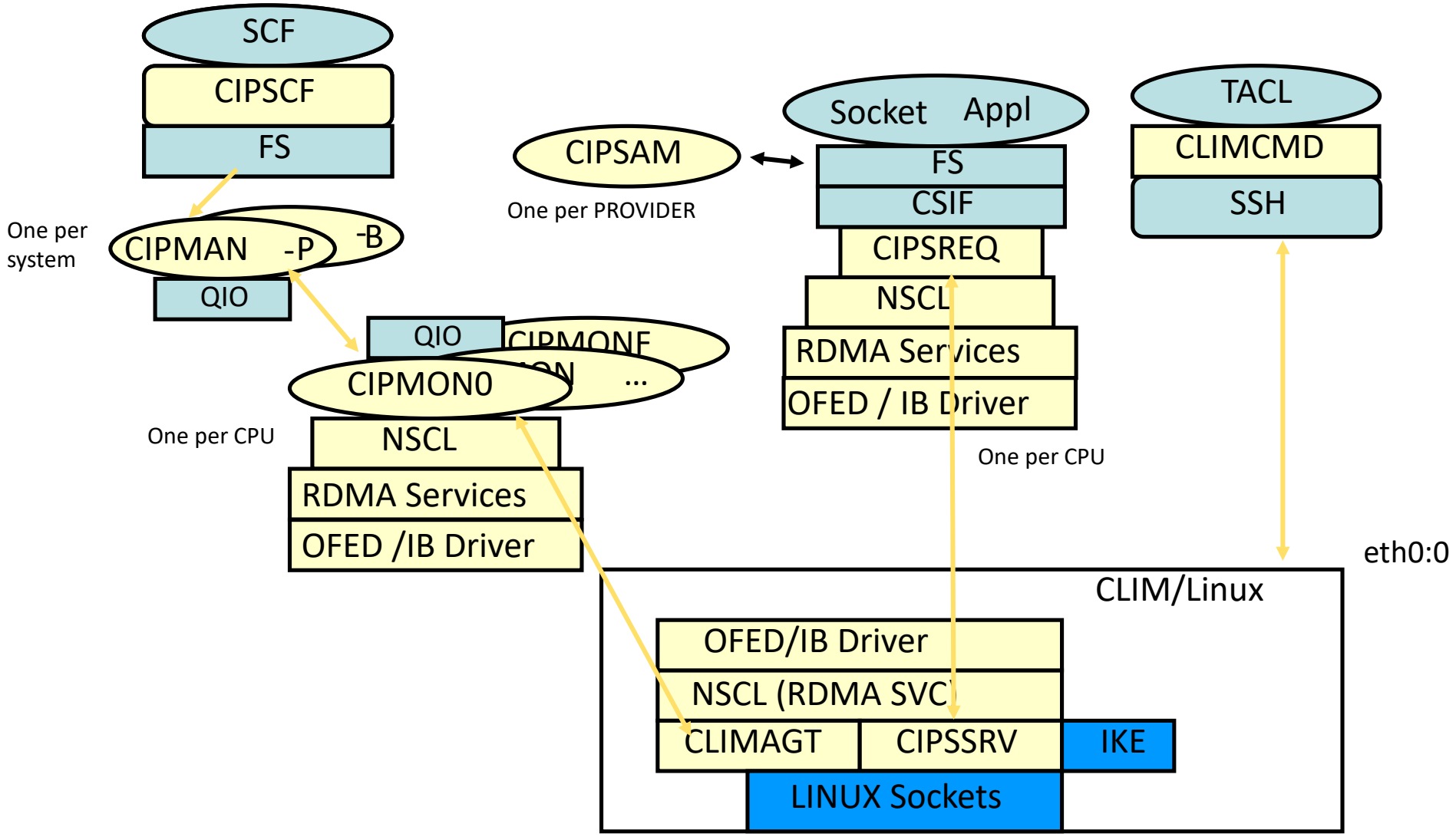
IB based TCP/IP Subsystem

Theory of Operation - IP CLIM

- Cluster I/O Protocol (CIP) Subsystem
  - Provides a configuration and management interface for I/O
- Uses a Linux server as a front end
  - Moves network stacks from the NonStop host to a Linux server
- Leverages the Open-Source Linux environment
- Provides support for IPsec, SCTP, IPv6
- Dual Infiniband fabrics connections
- Provides Gigabit and 10 Gigabit Ethernet interfaces
- The CLIM provides the physical interface to the network or storage devices



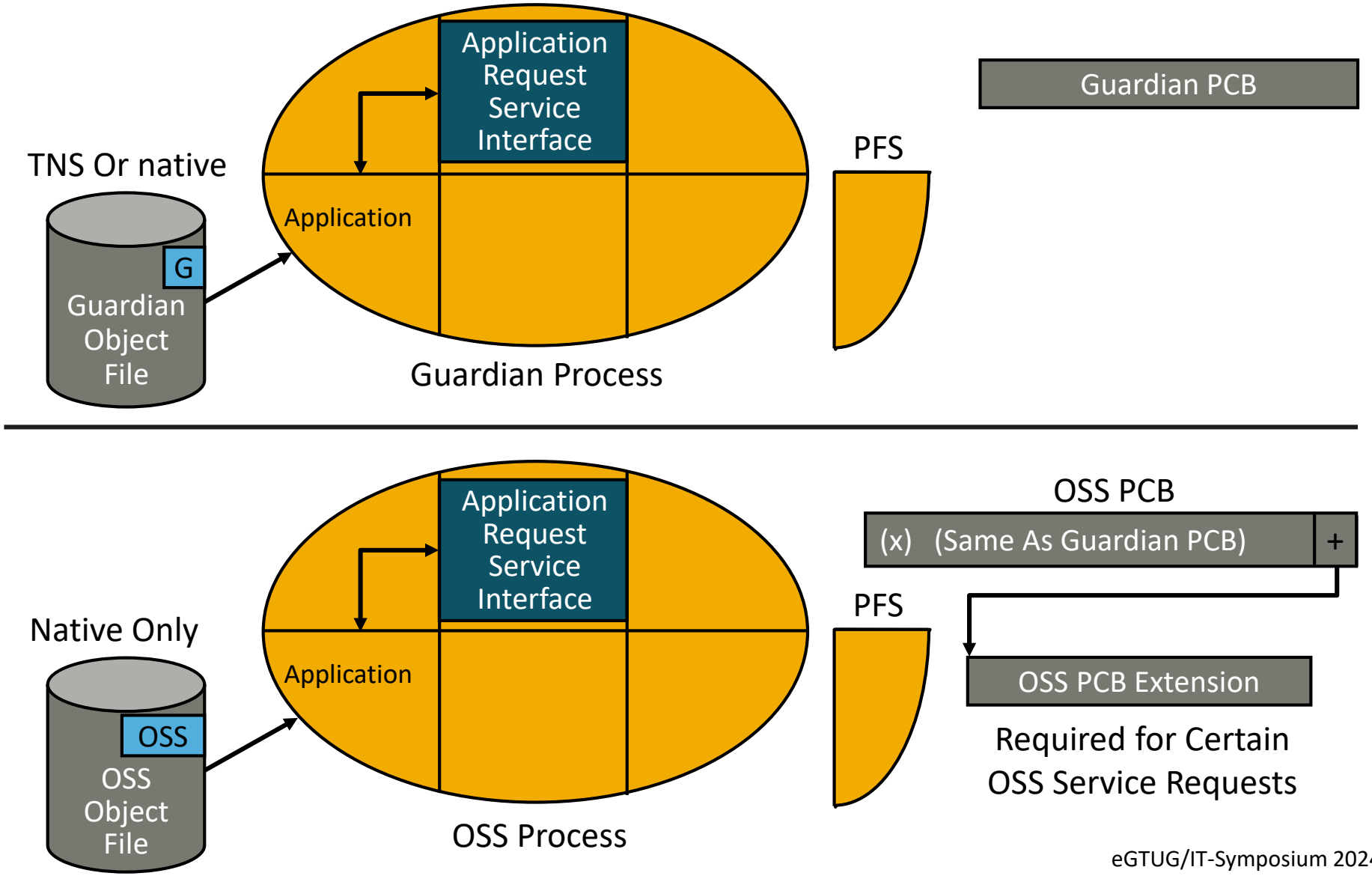
CIP Software Architecture Overview (IP CLIM (XPF version))



Part 4

- Debugging
- Run-Time architecture
  - TNS Prtocesses
  - TNS/X processes
- Process control
  - Guardian
  - OSS

Guardian Versus OSS Processes

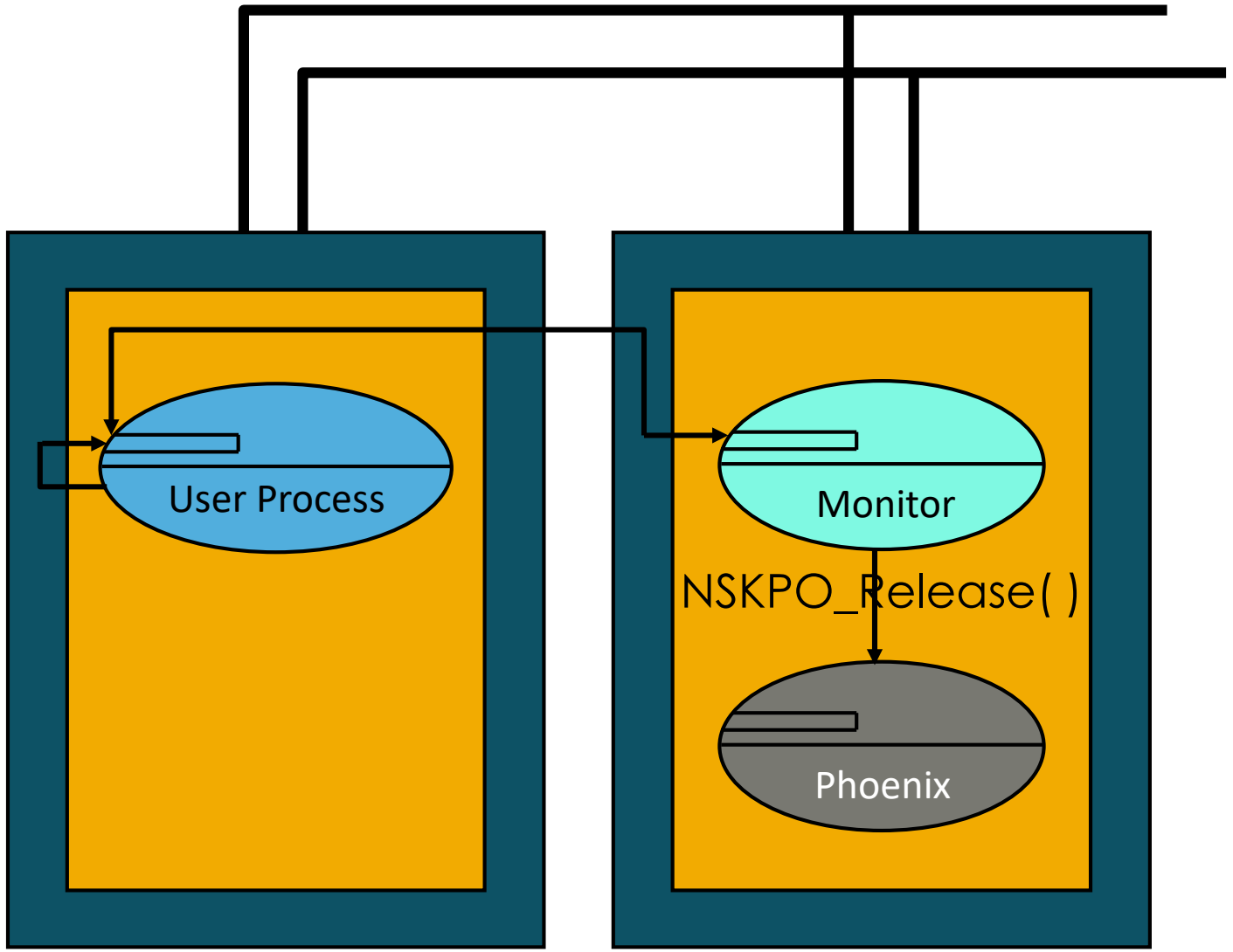


Guardian Process Control

Process Creation

- Guardian environment:
  - NEWPROCESS
    - Uses PID for the process identifier.
    - PID cannot identify process numbers > 255.
  - PROCESS_CREATE_
    - Uses PHANDLE (process handle) to identify processes.
  - PROCESS_LAUNCH_
    - Can override native process defaults.
    - Uses an extensible struct to pass parameters.
  - From a TACL prompt, the RUN command uses PROCESS_LAUNCH_.

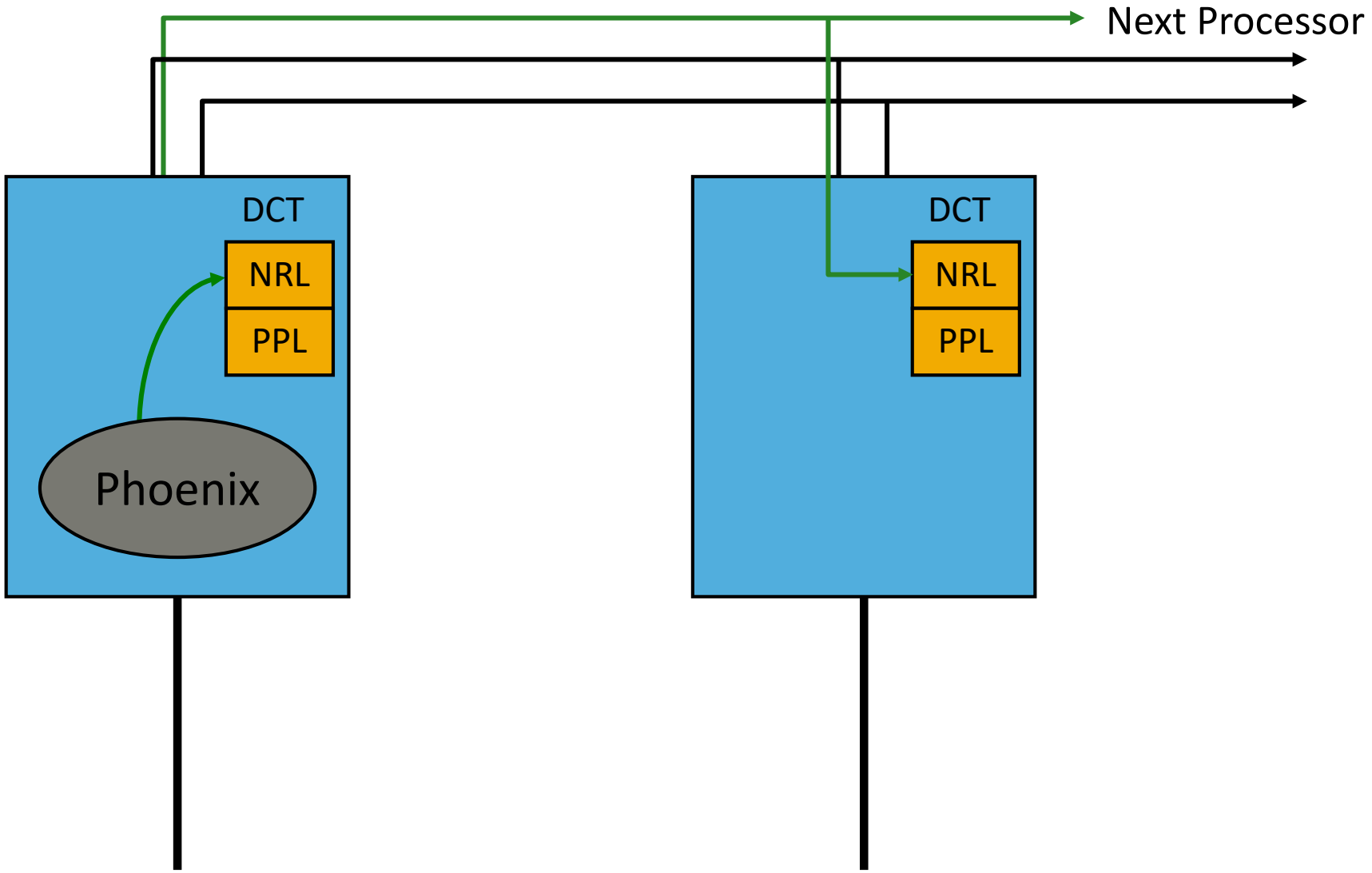
Process Creation — Guardian Environment



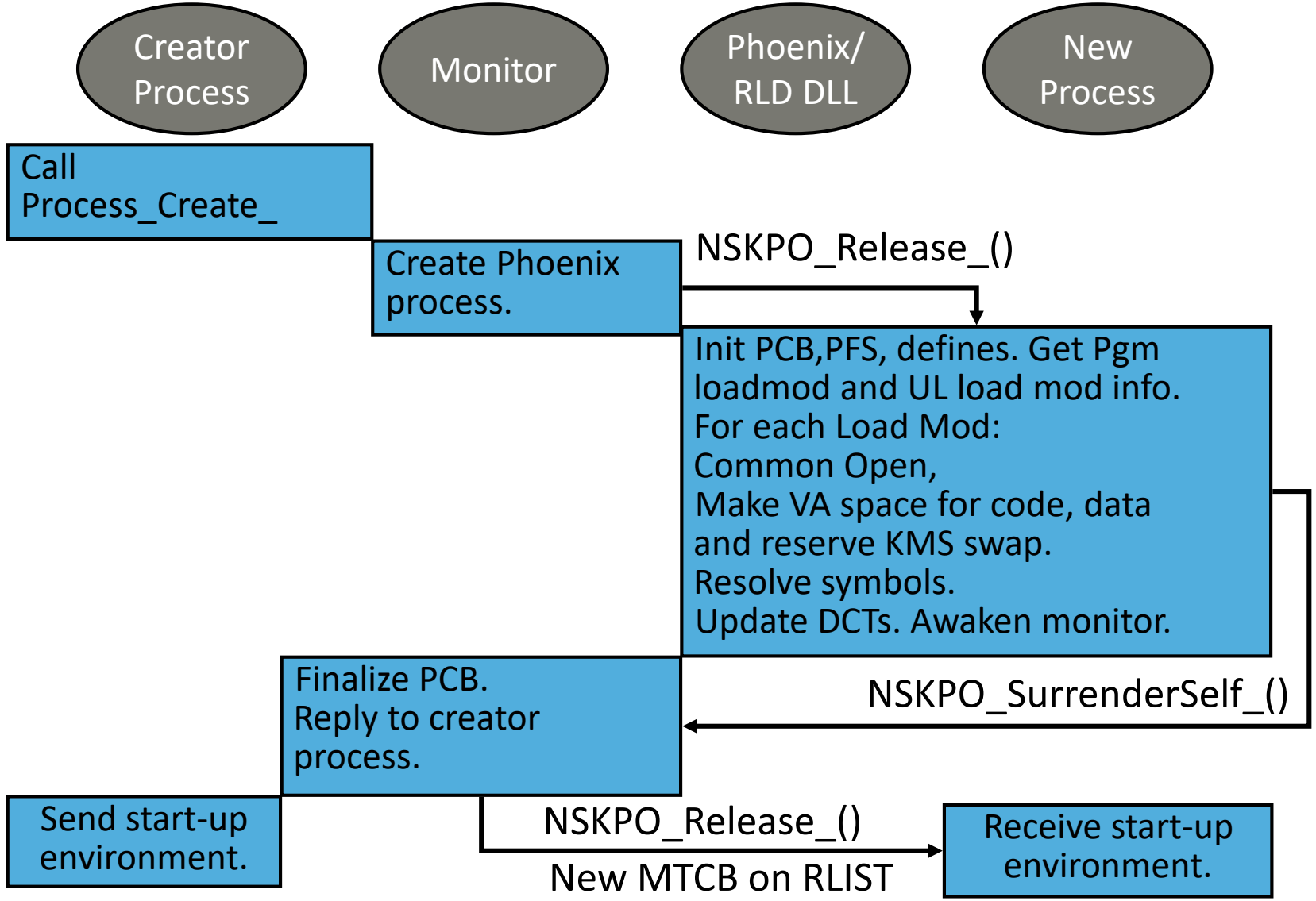
RLD Symbol Resolution

- For each load module:
  - Determines what symbols each load module defines.
  - Determines what symbols each load module needs resolved.
  - Resolves each symbol from the loadlist.
  - Fills in the GOT tables with the values of the symbols.

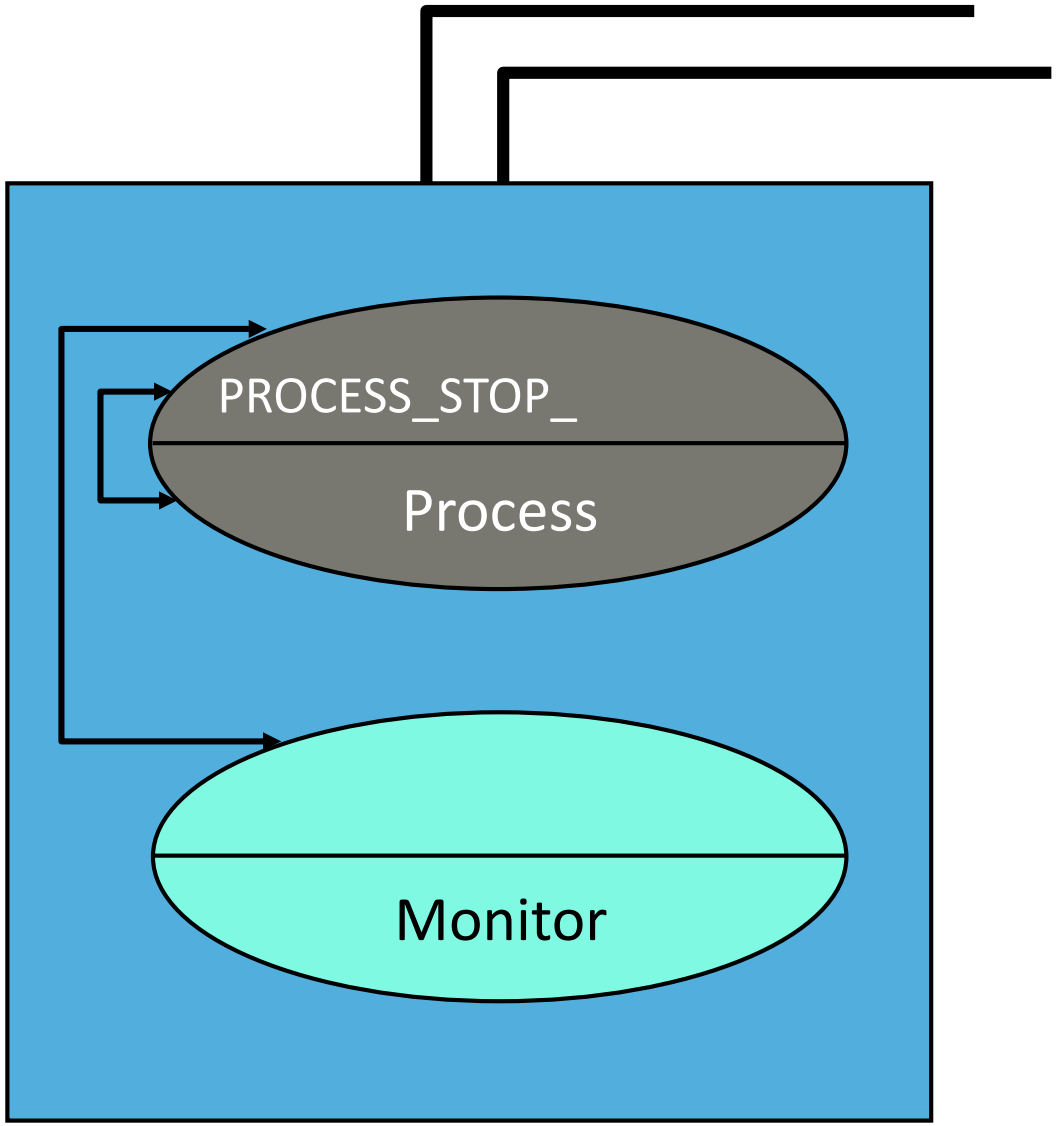
Phoenix Handling of Named Processes



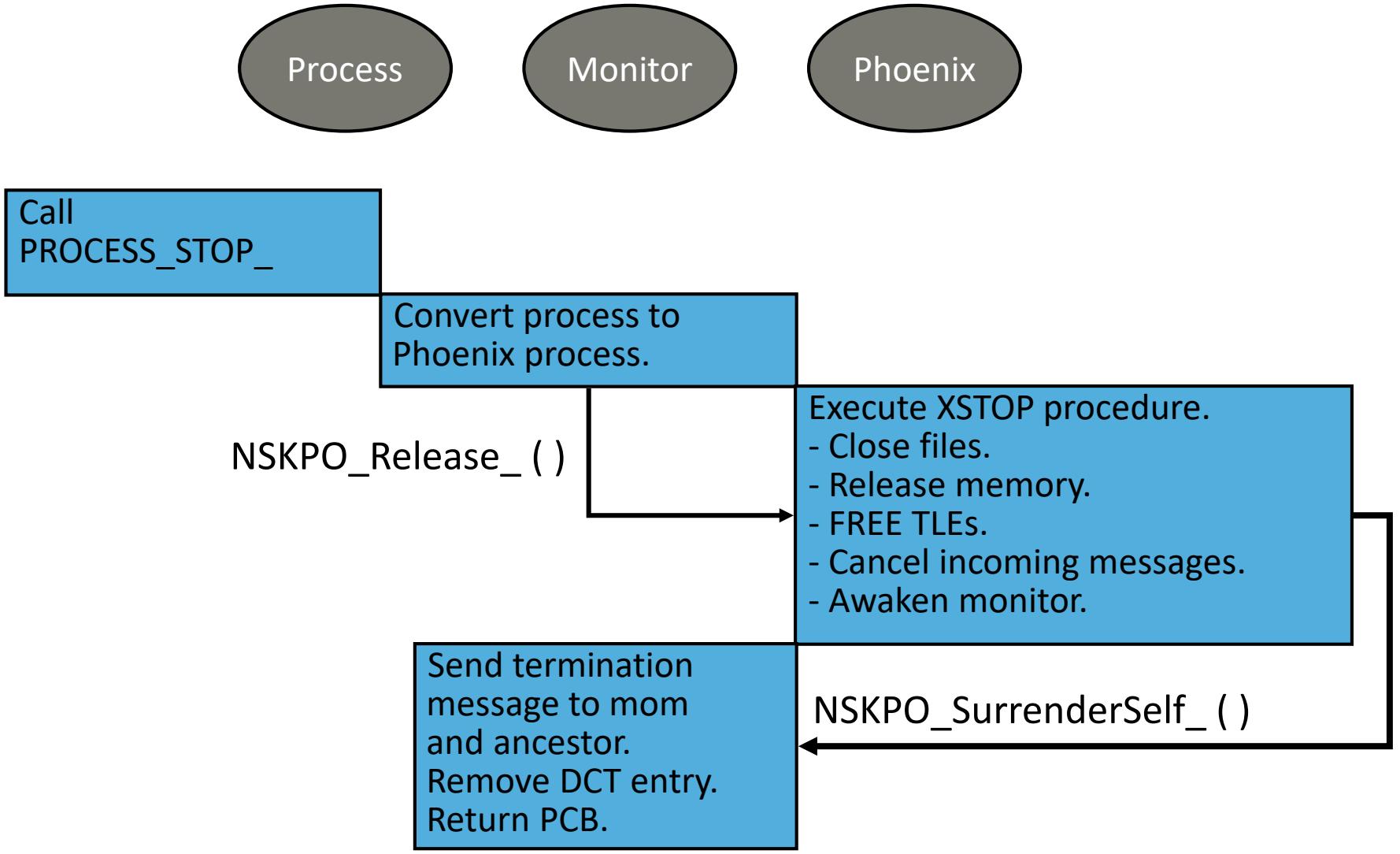
Guardian Process Creation — Summary



Process Termination —PROCESS_STOP_ Procedure



Guardian Termination — Summary



Monitor Functions

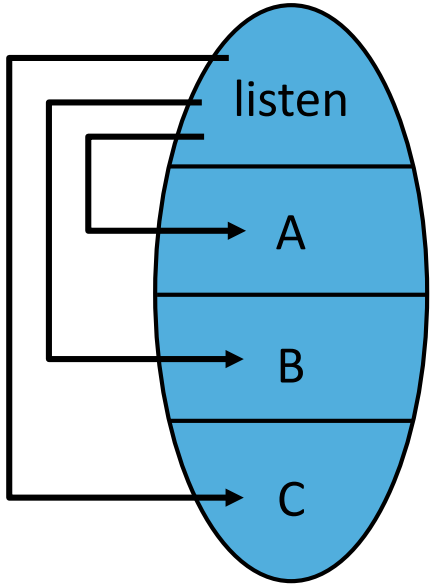
- Starts and stops processes.
  - STM and others give Monitor the process to terminate after fatal traps.
  - Dynamic system configuration.
- Performs process control.
- Returns information.
- Maintains the system time-of-day.

OSS Process Control

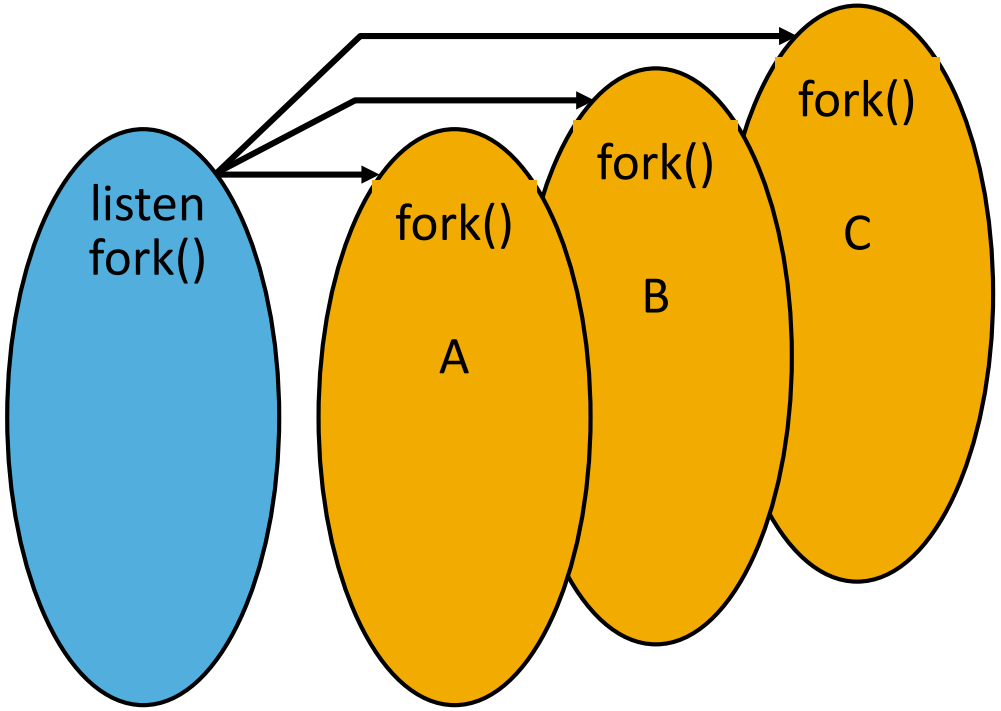
OSS Process Creation

- OSS environment.
  - fork().
  - exec*().
  - Shell run.
  - tdm_spawn().
  - tdm_*() takes additional parameters similar to PROCESS_LAUNCH_.
  - PROCESS_SPAWN_ - can be used from a Guardian process.
    - It is used by the OSH program.

Process Creation — OSS

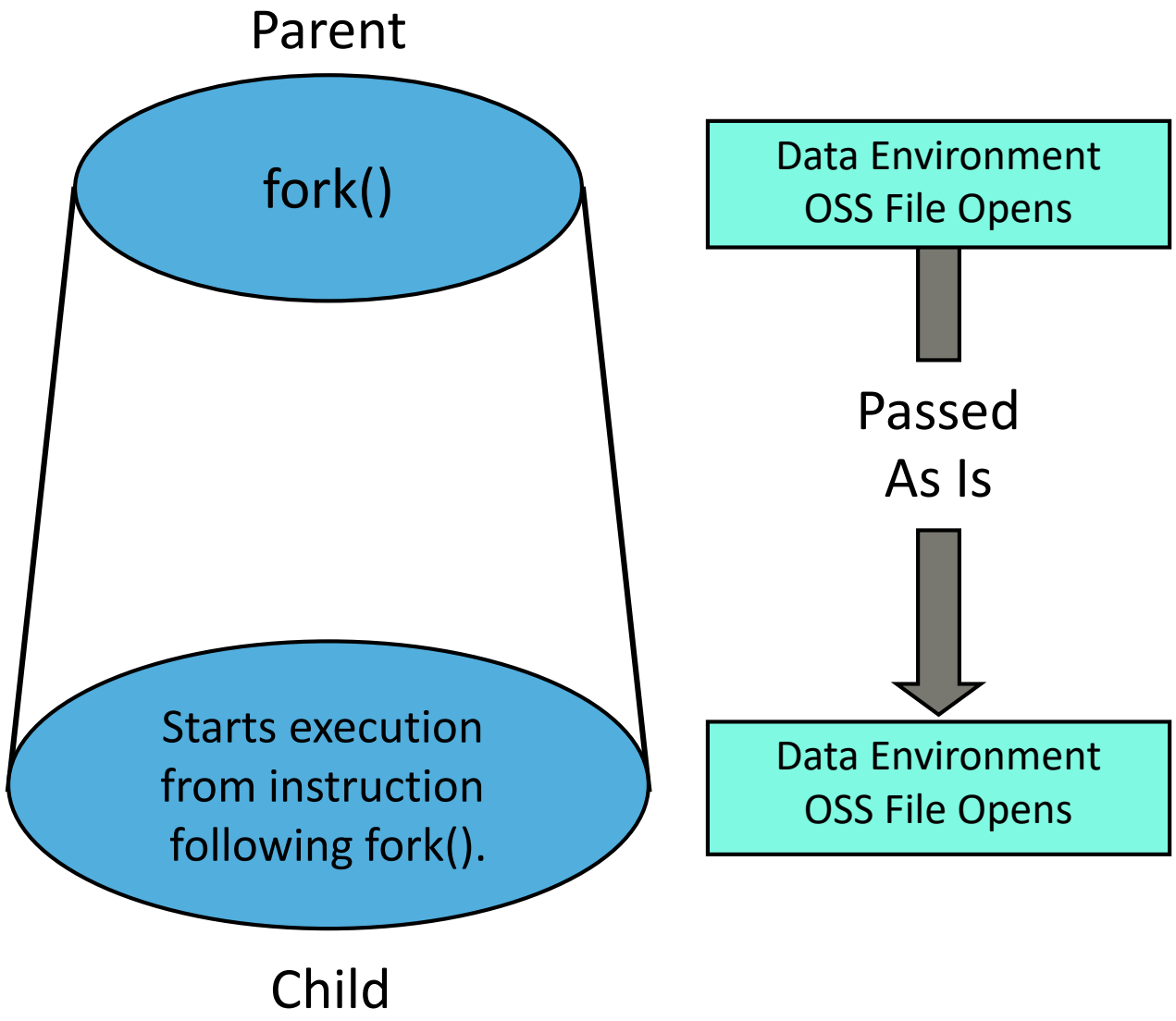


1 — Single, monolithic program, multithreaded, difficult to write and to test.
Each thread handles one request.

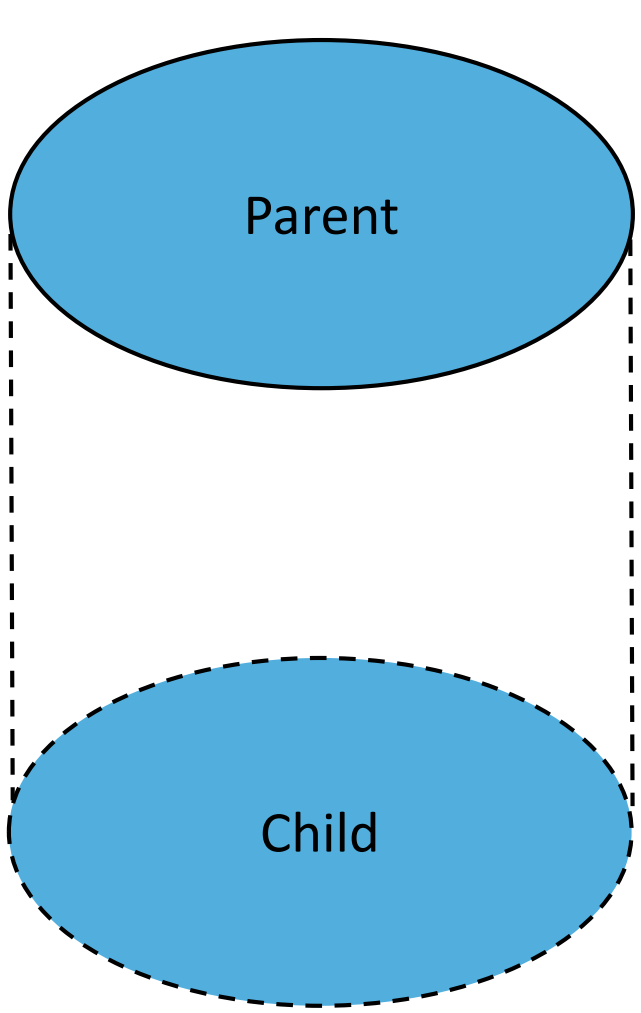


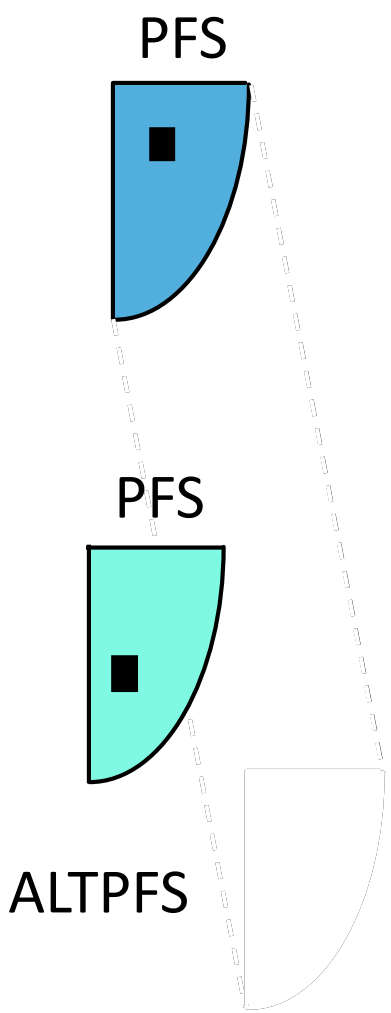
2 — One program forks itself, keeping its opens, and each forked copy shares the code. Simpler to write and test because not multithreaded.
Each process handles one request.

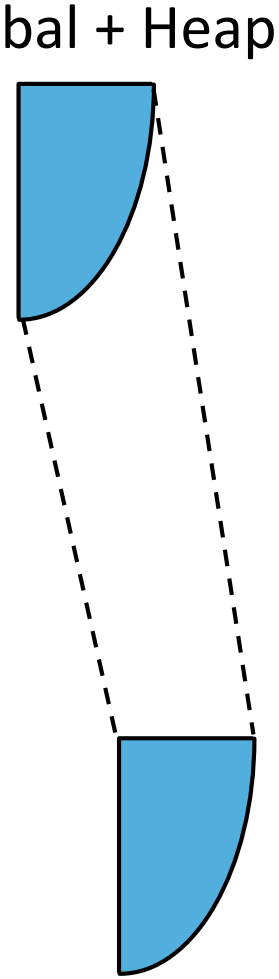
Process Creation — OSS fork()



The Copy Environments

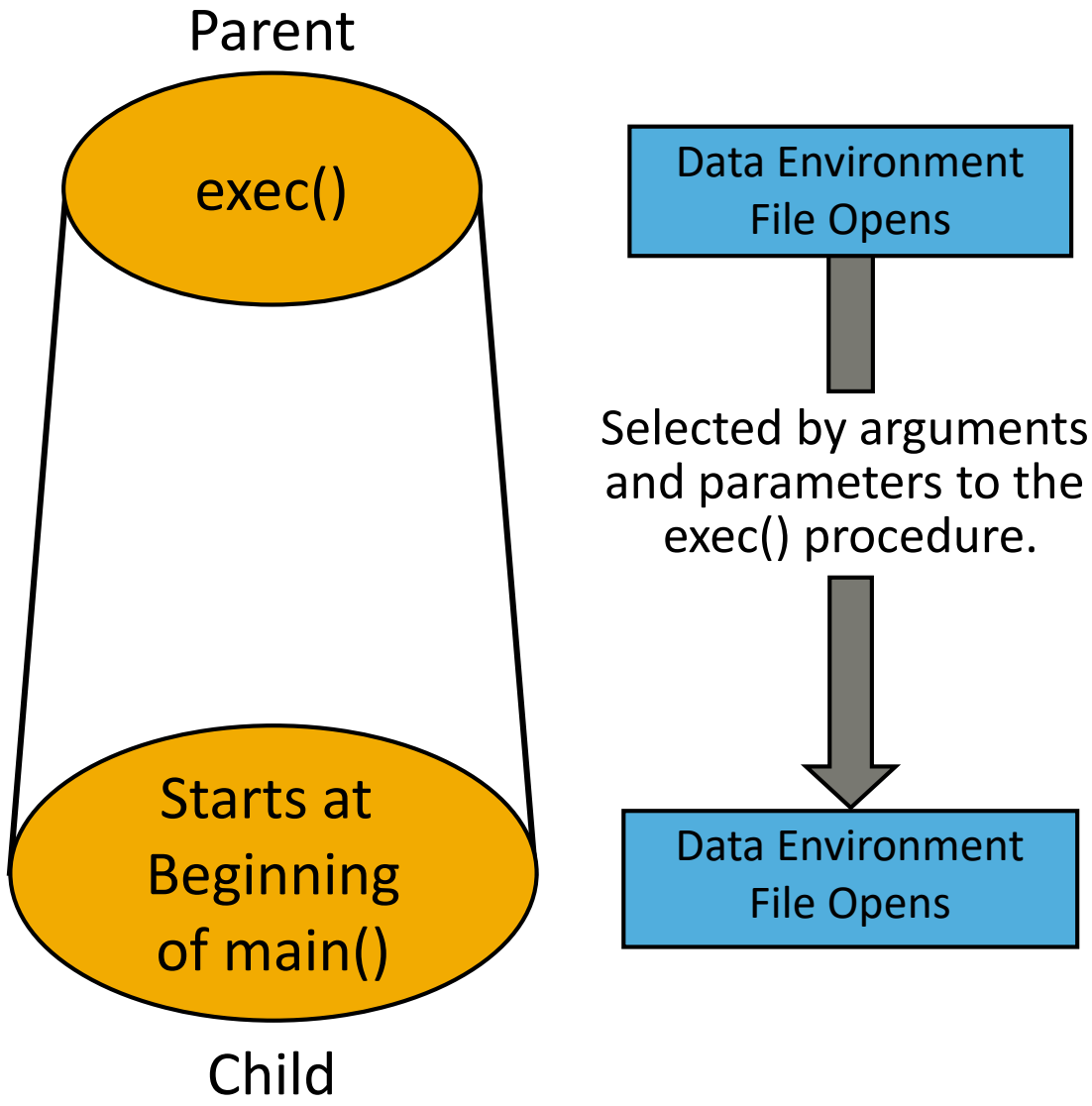




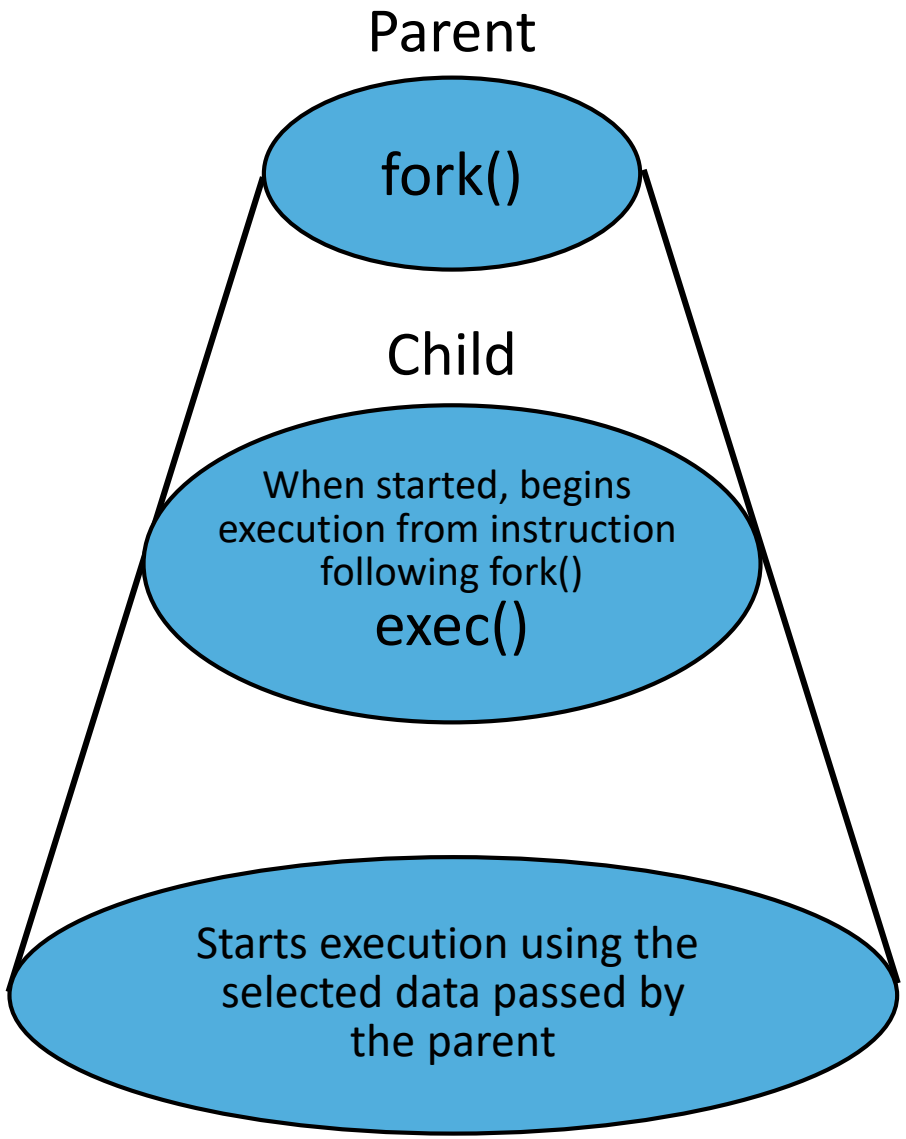


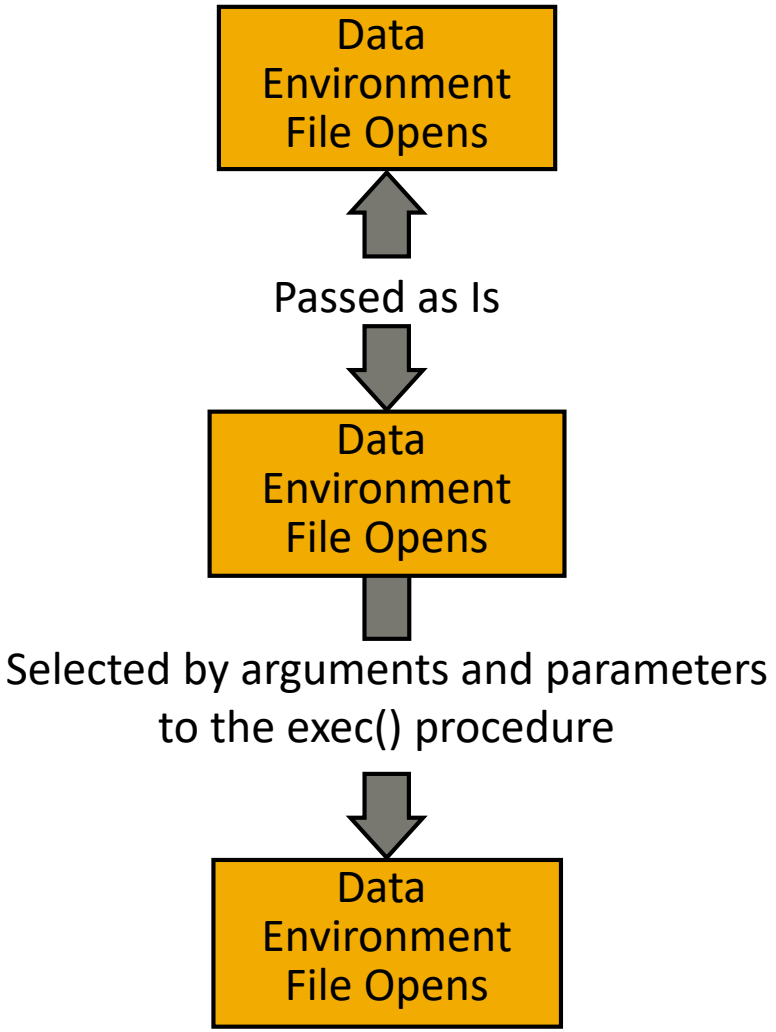
Process Creation — OSS exec()

Following the start of the child process, the parent goes away.

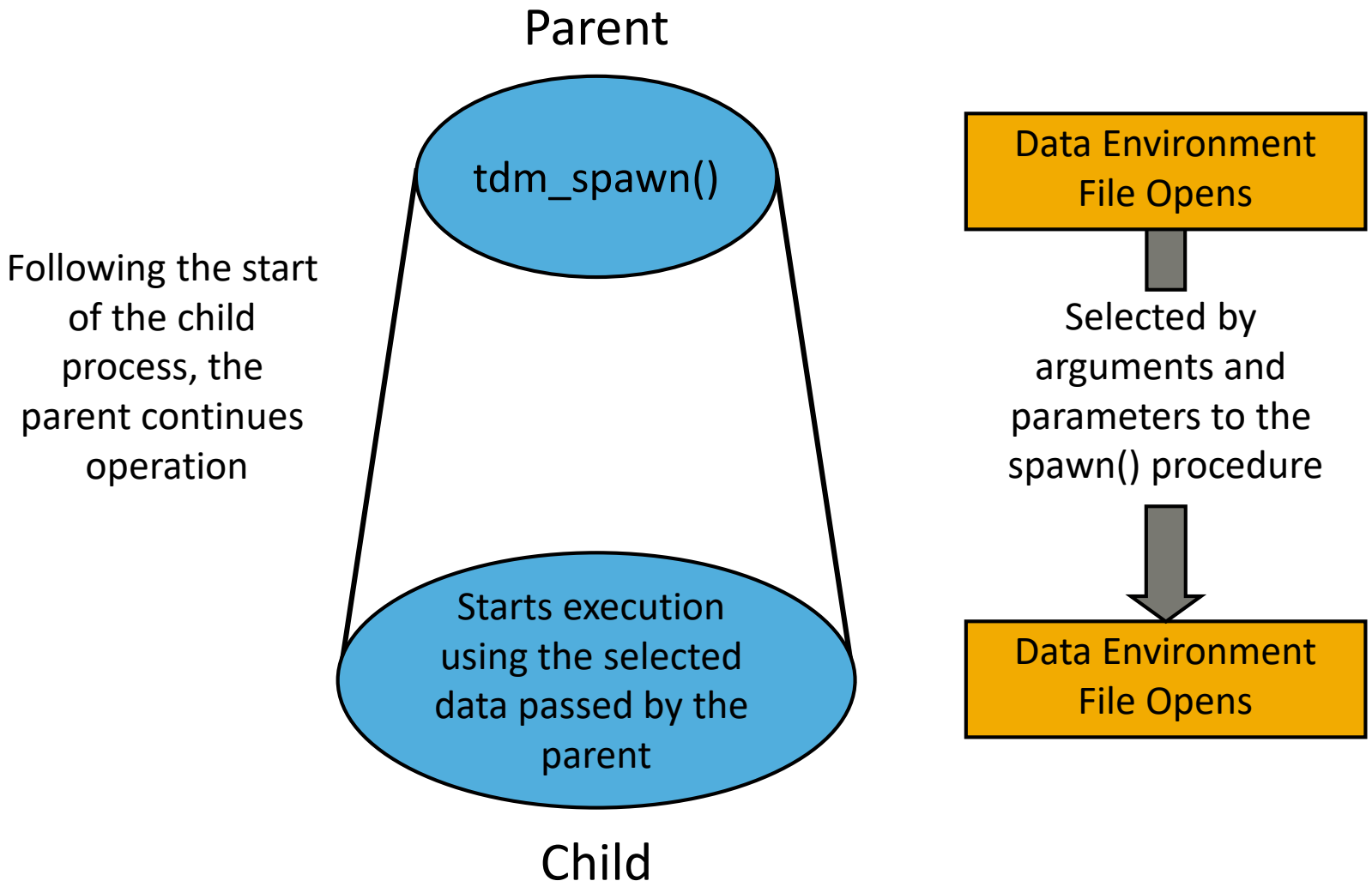


Process Creation — OSS fork(), exec()





Process Creation — OSS tdm_spawn()

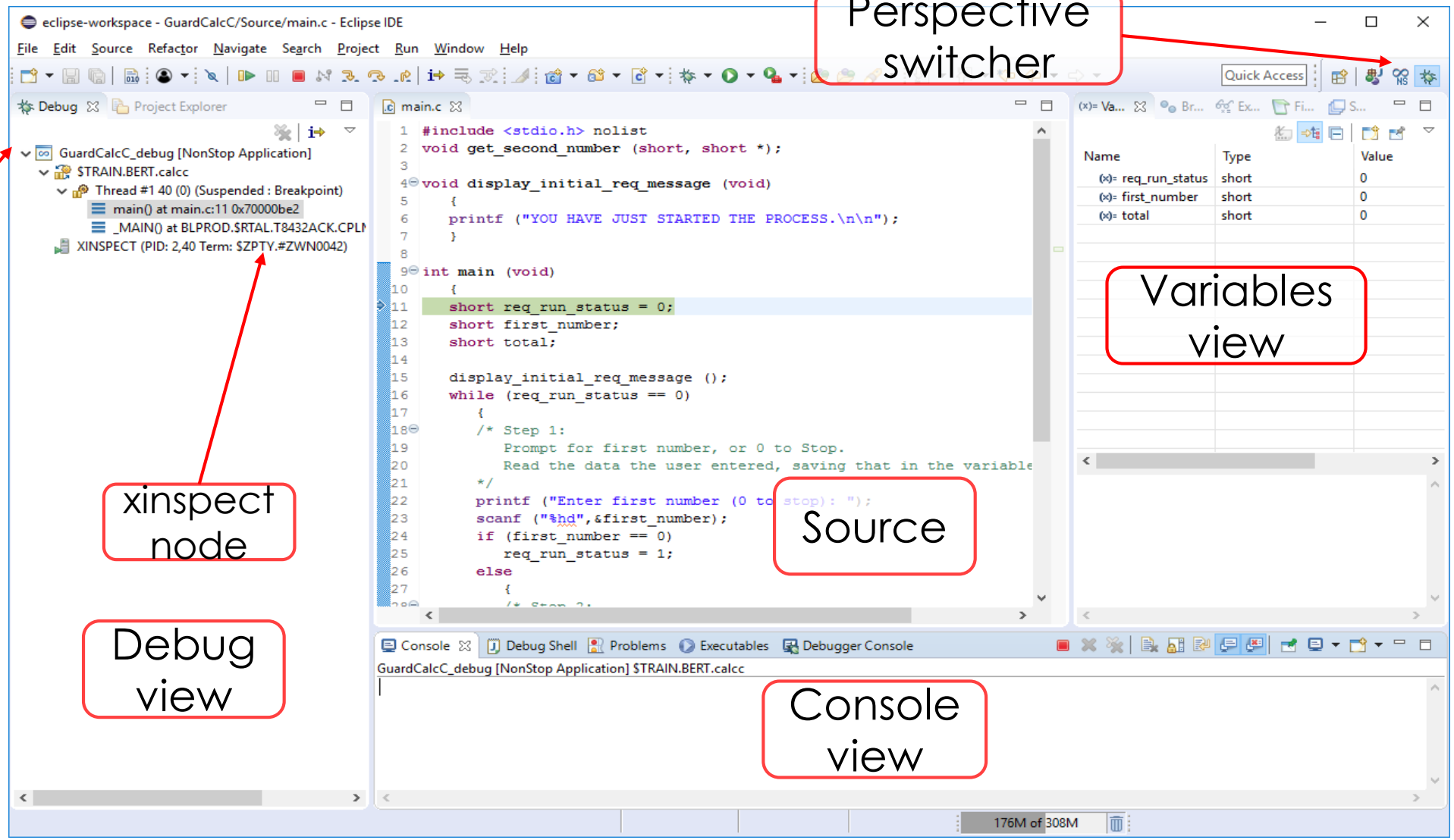


Debuggers

What are the L-Series Debuggers?

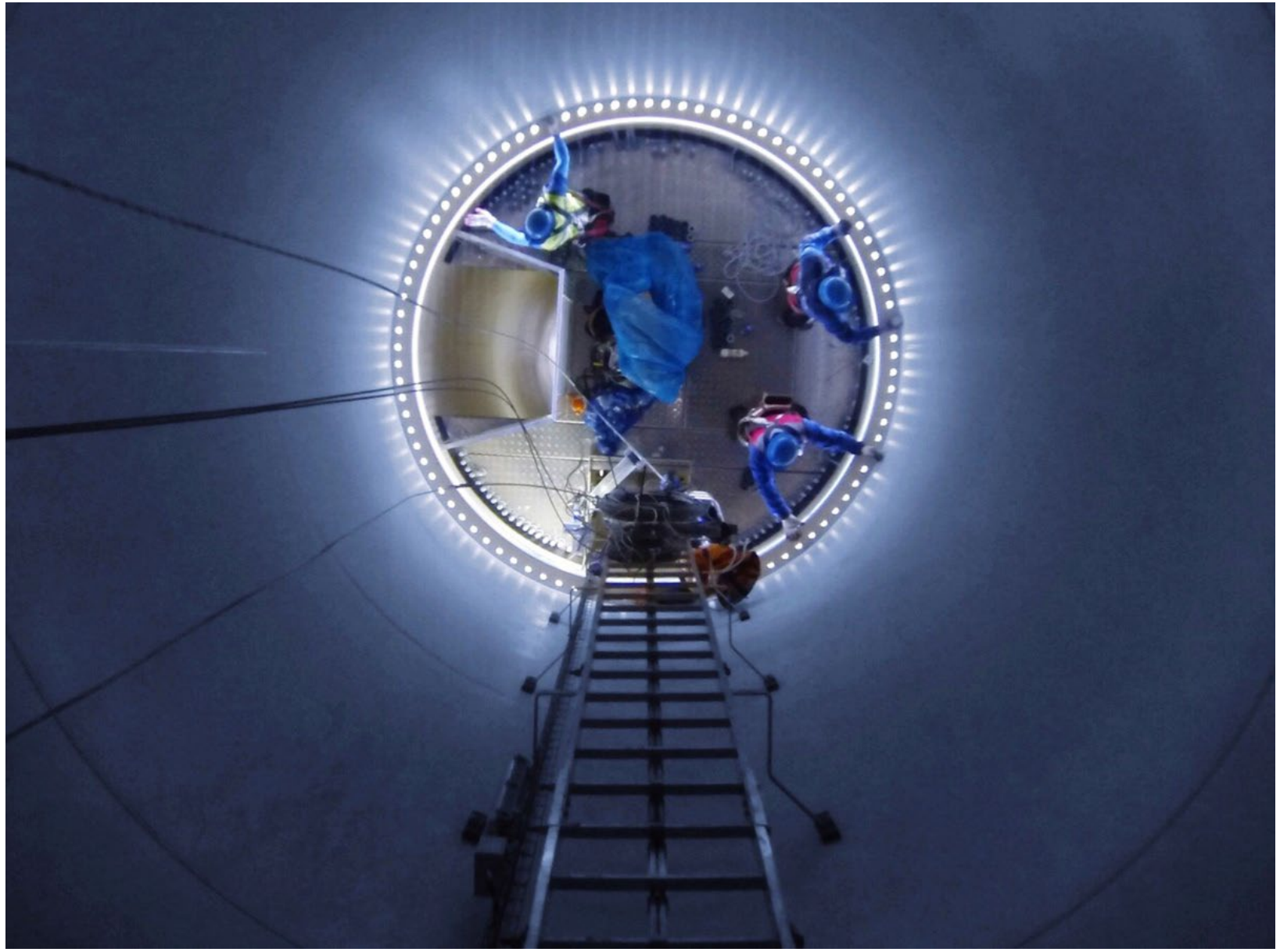
- xInspect (Native Inspect):
  - Original based on GNU debugger and running from the TNS/X command line.
- NSDEE (Eclipse) version 13:
  - Supports Native TNS/E and TNS/X debugging (no TNS)
  - Part of Eclipse development environment
  - PC hosted GUI
- Inspect:
  - TNS debugging only.
- TNS Visual Debugger (TNSVDBG):
  - TNS debugging only
  - Basically, Visual Inspect but restricted to TNS only debugging on TNS/X systems

Debug Perspective



Part 5

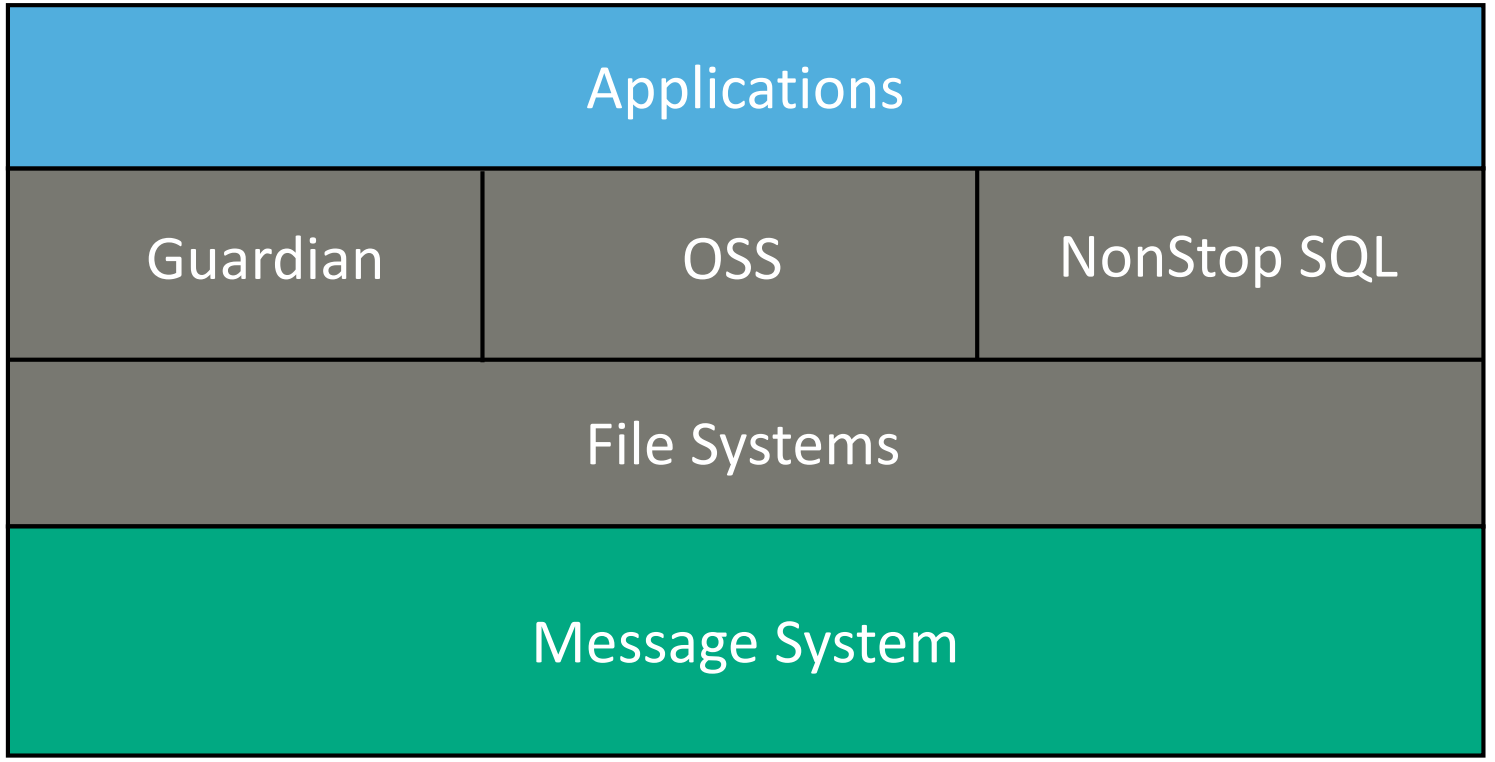
- Guardian file system
- Open System Services file system



Functions of the File System

- Callable interface to message system.
- Logical file name to process handle (phandle) resolution.
- Passes security information to the IOP or file-system layer of the application server.
- Device independence: Everything is treated as a file.
- Keeps track of outstanding messages and associated buffers.
- Some fault tolerance support.
- (For disk) Partition file and alternate-key support.
- What the file system does not do: IOPs and TMF software, though not part of the file system, provide functions that are often associated with a file system.

File Systems

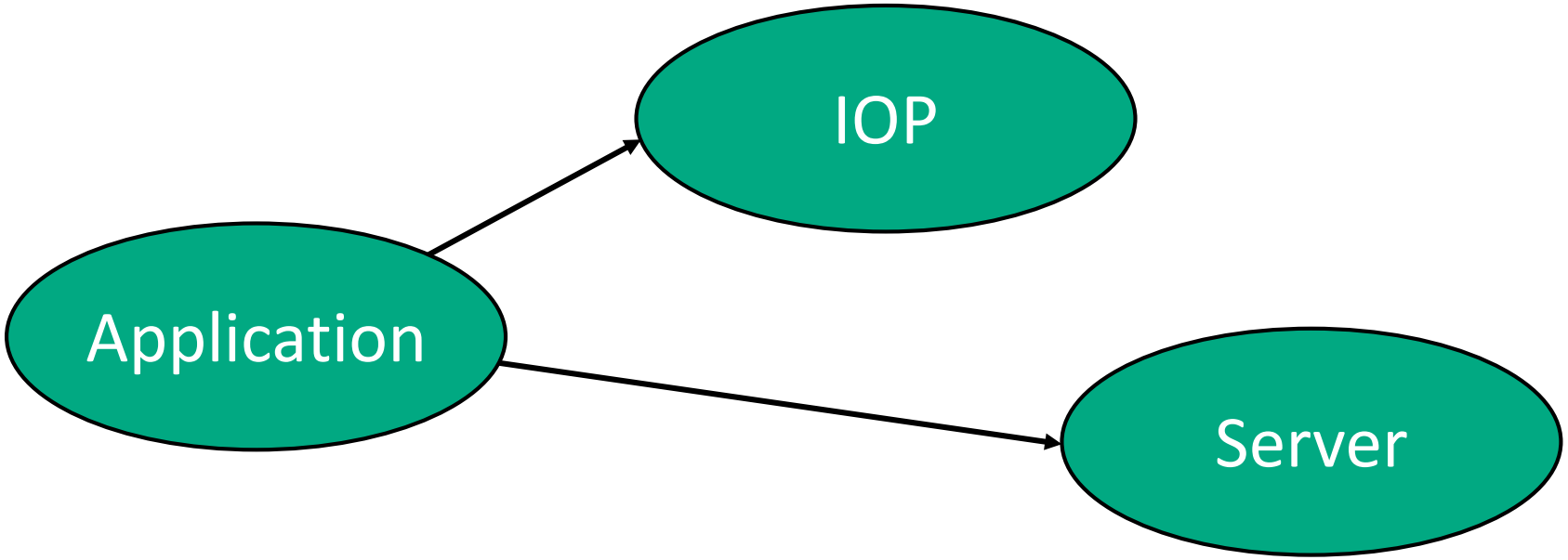


Control Structures for Guardian File System

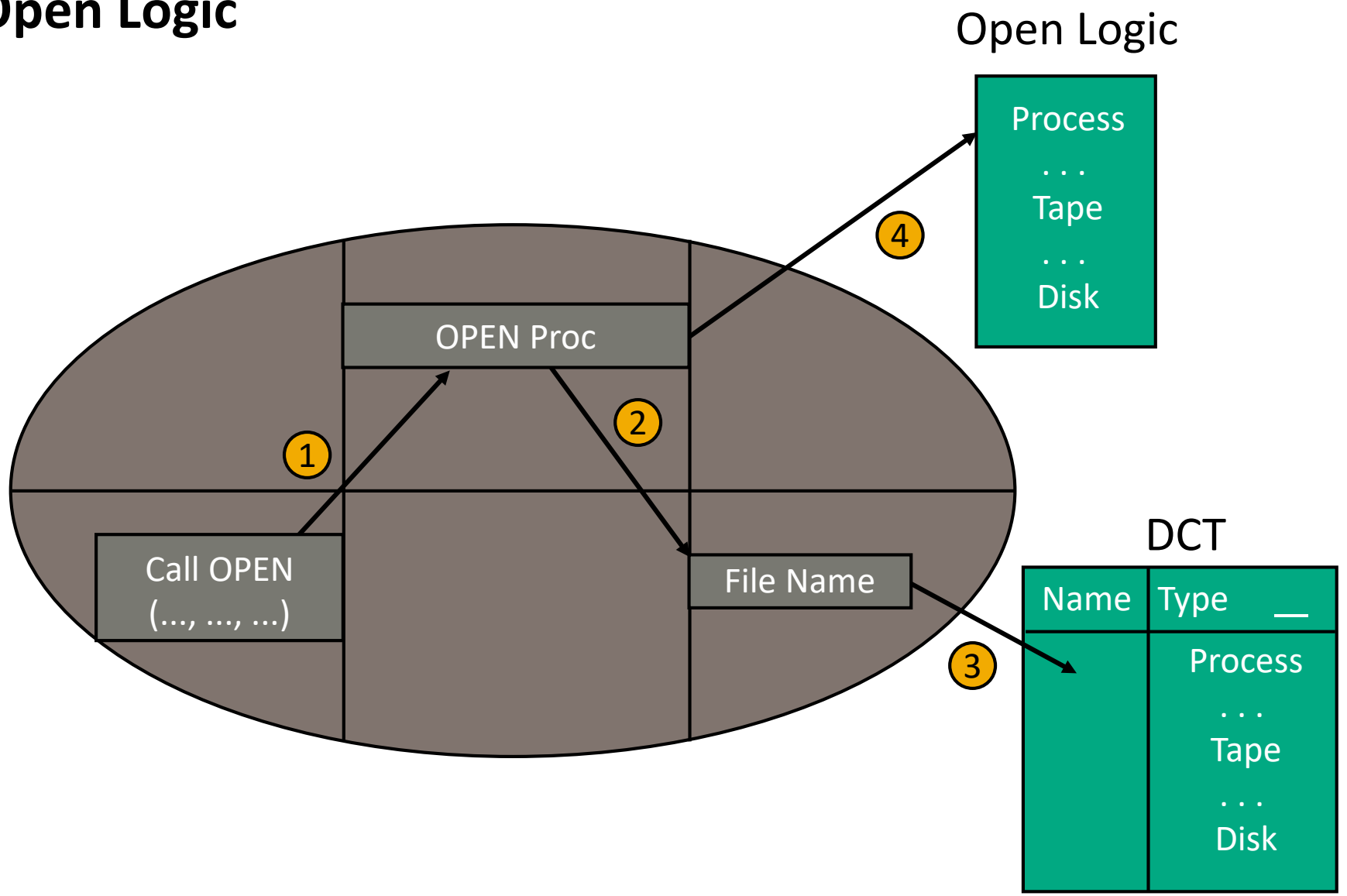
- Access control block (ACB) — Opener controlled.
- Destination control table (DCT) — Kernel-reserved segments.
- File table.
- IOP has open control block (OCB) corresponding to ACB.
- Disk process has a file control block (FCB) to represent each open disk file.

Open Request Processing Between Requester and Server

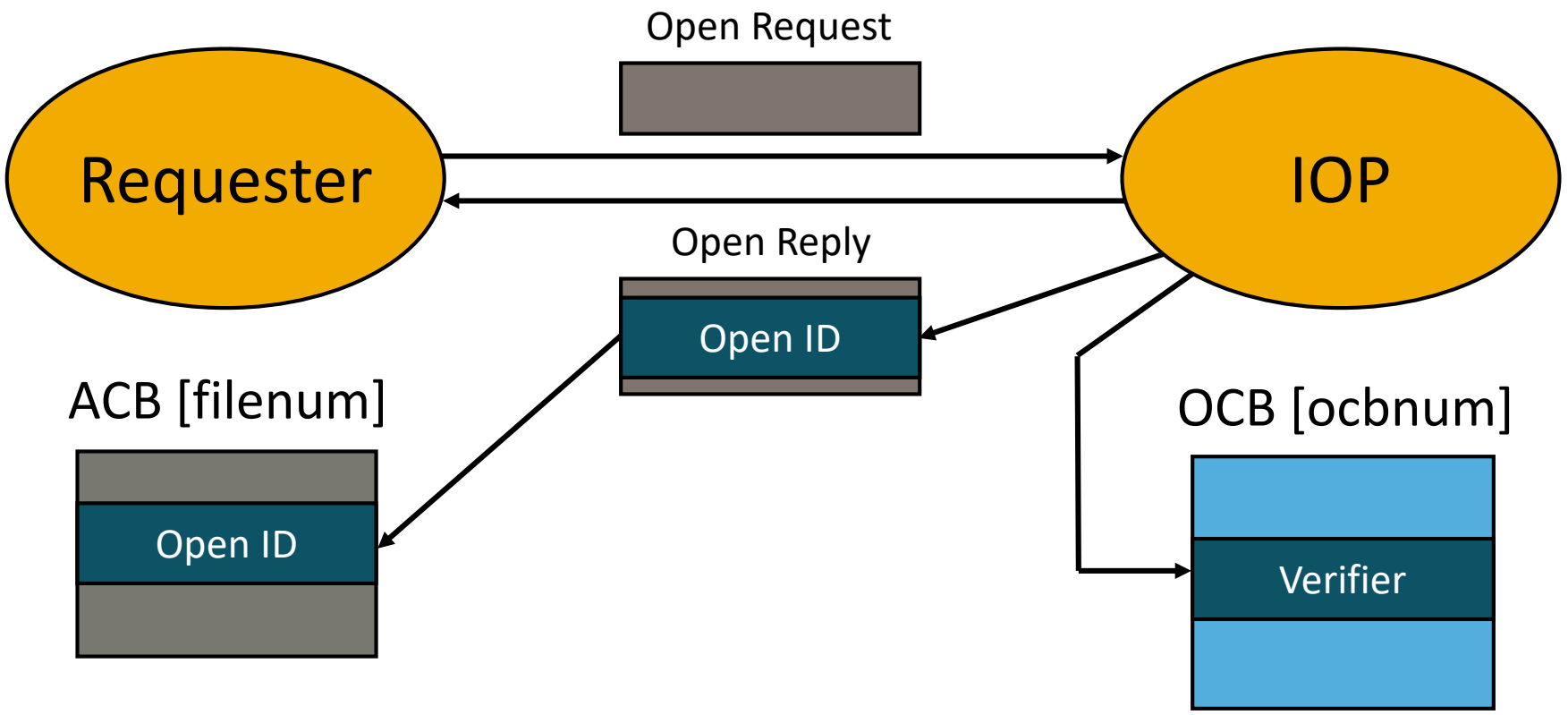
Application Might Send OPEN	Case 1 to IOP	Case 2 Or To Application



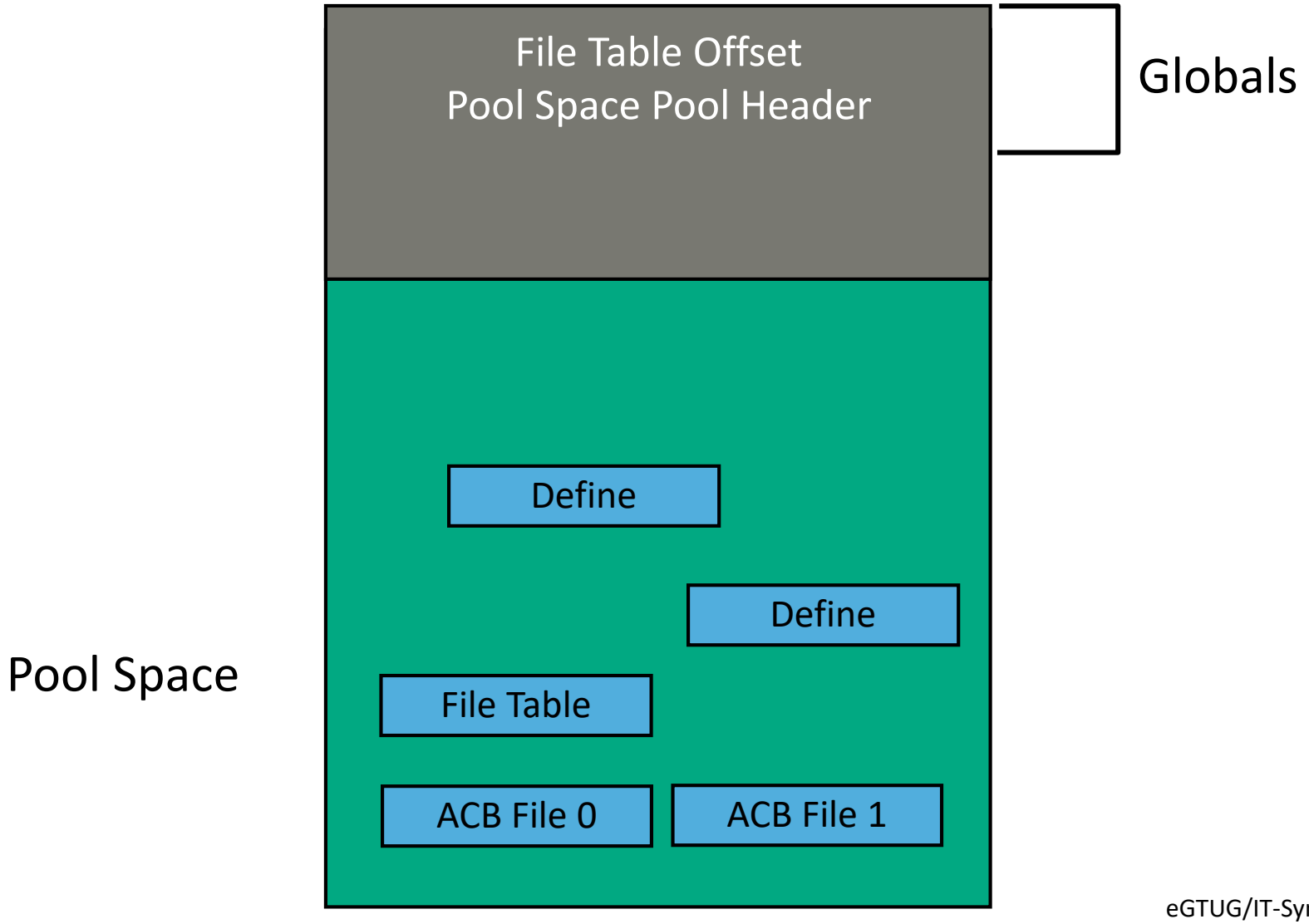
Requester Open Logic



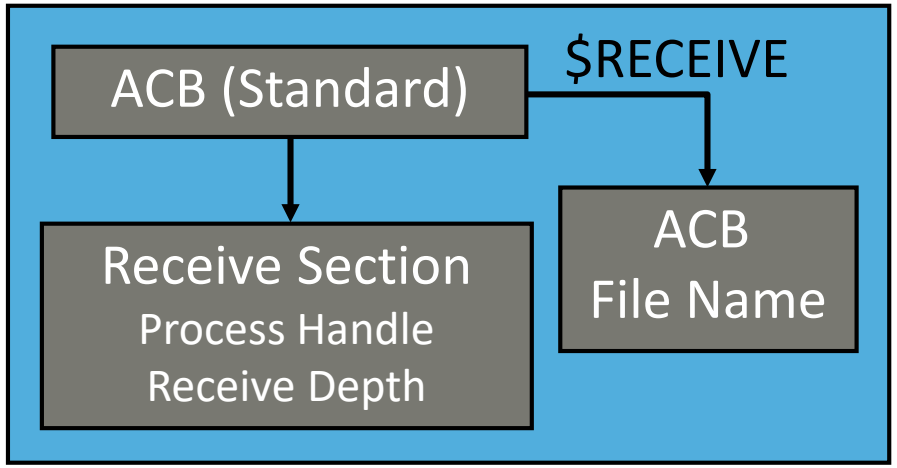
Open Request Processing By the IOP

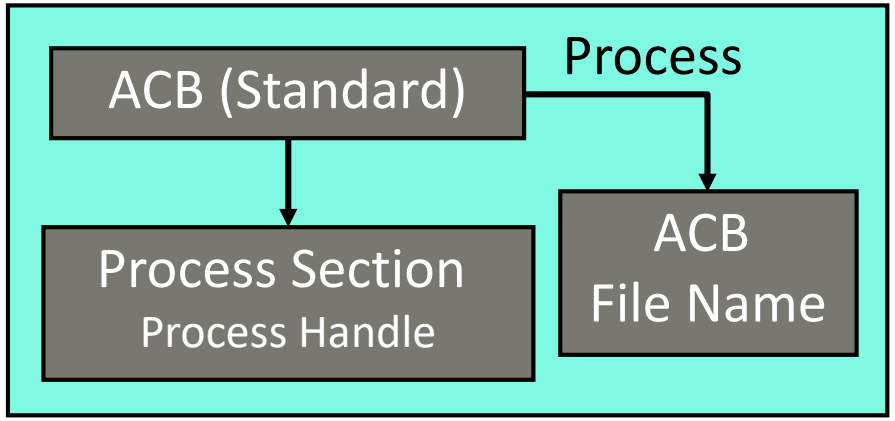


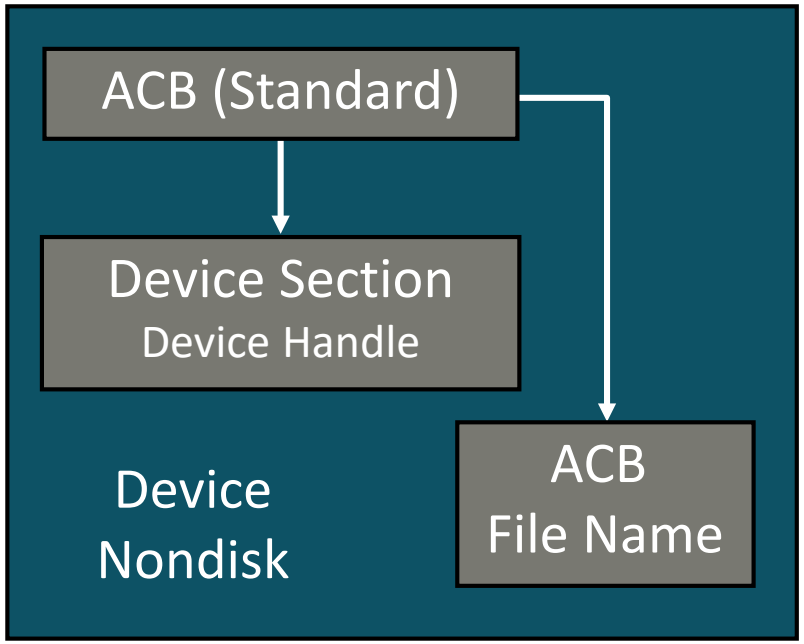
Process-File Segment (PFS)



ACBs, Guardian File System





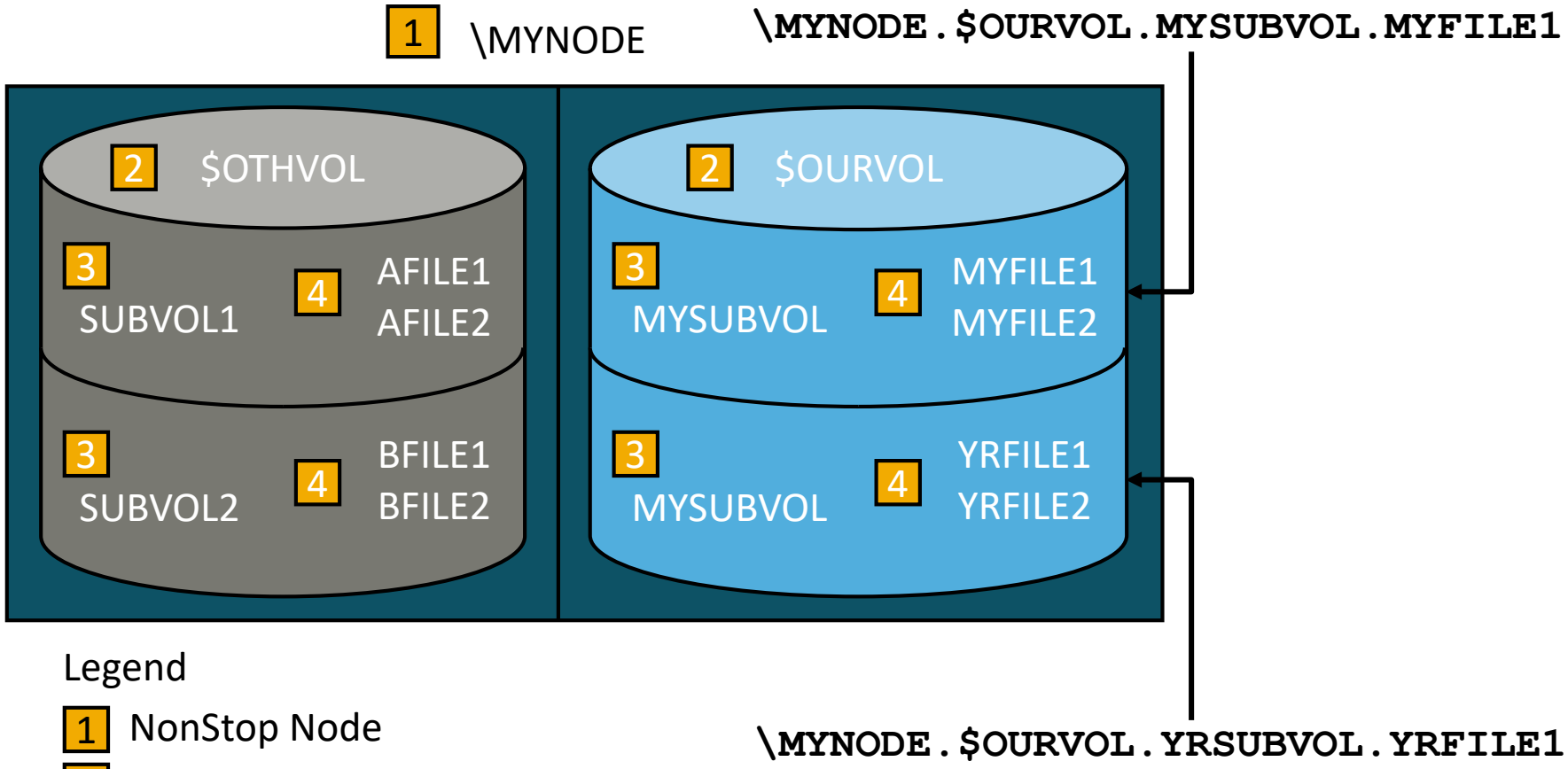


OSS File System

- OSS file system refers to an entire collection of files.
- File system includes a hierarchical structure of directories, subdirectories, and files.
- File system has a single root.

Guardian Disk File-Name Hierarchy Has Few Levels

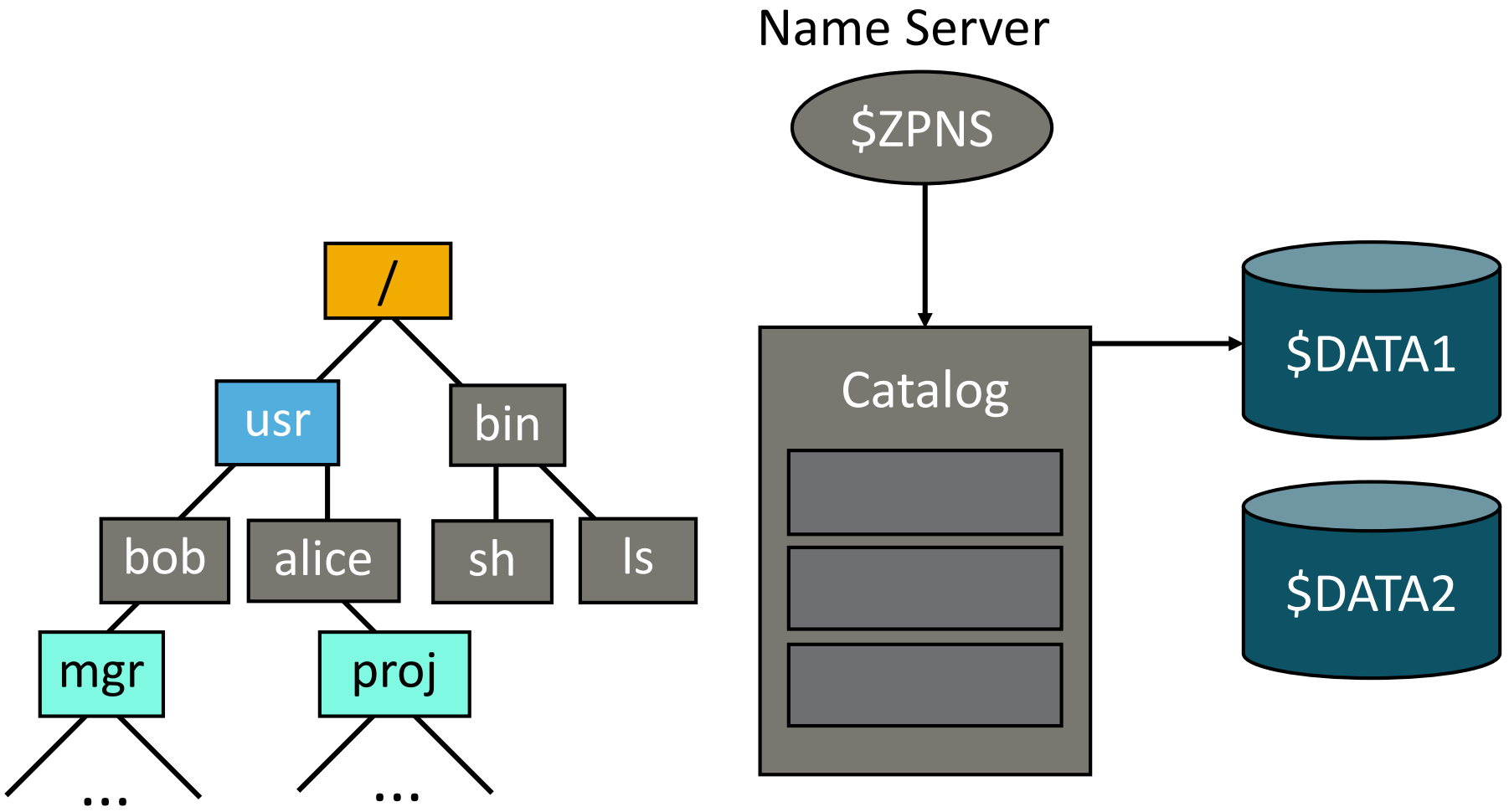
- Guardian Disk File Organization



Legend

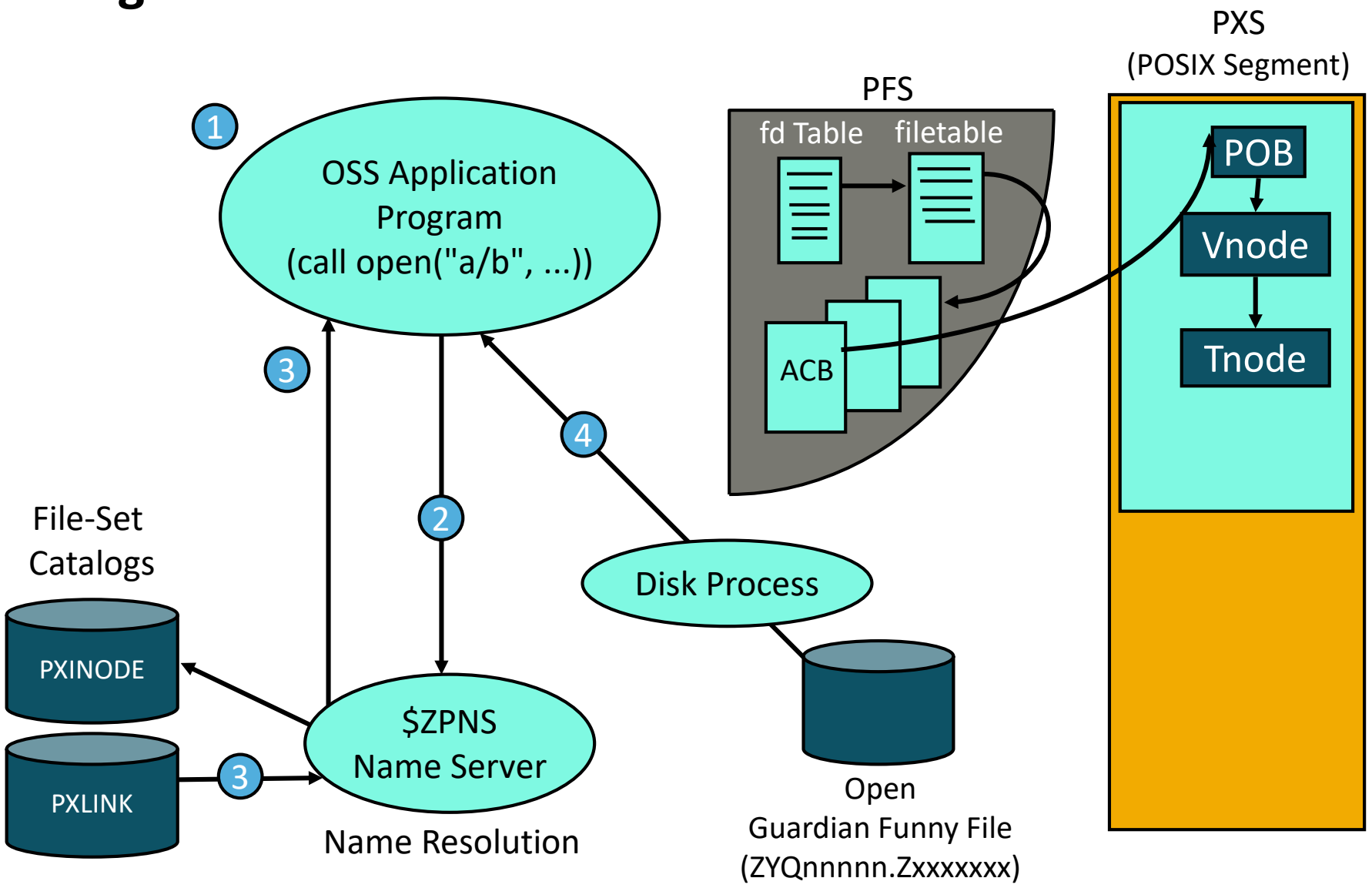
1 NonStop Node
2 Disks
3 Subvolumes
4 File ID

OSS Pathname Mapping



`/usr/alice/proj/schedule` → `$DATA1.ZYQ00003.Z000002A`

OSS Open — Regular

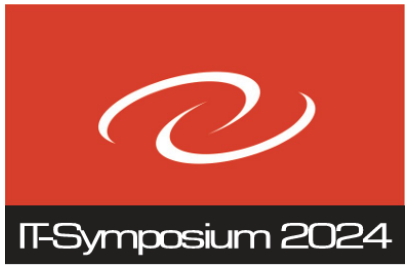


Would you like to know more?

- This slide deck is a heavy compressed version of the 5 day "**HPE Integrity NonStop Operating System Architecture U8609S**" course.
- If you would like to know more, you should enrol for the U8609S course.
- The U8609S course is scheduled in the EMEA (CET) and USA (CT) time zone and delivered in English.
- Current scheduled sessions:
  - **May 6**, US Central Time
  - **June 17**, Central European Time
  - **October 28**, US Central Time
  - **November 18**, Central European Time
- Check out any changes in our schedule and register on: www.nonstop-academy.com

THANK YOU

Questions?



Thank you for attending this talk
What is the NonStop Crown Jewel?

Bert van Es
bert.van.es@continuous.nl

NonStop Academy