IPADS
INSTITUTE OF PARALLEL AND DISTRIBUTED SYSTEMS

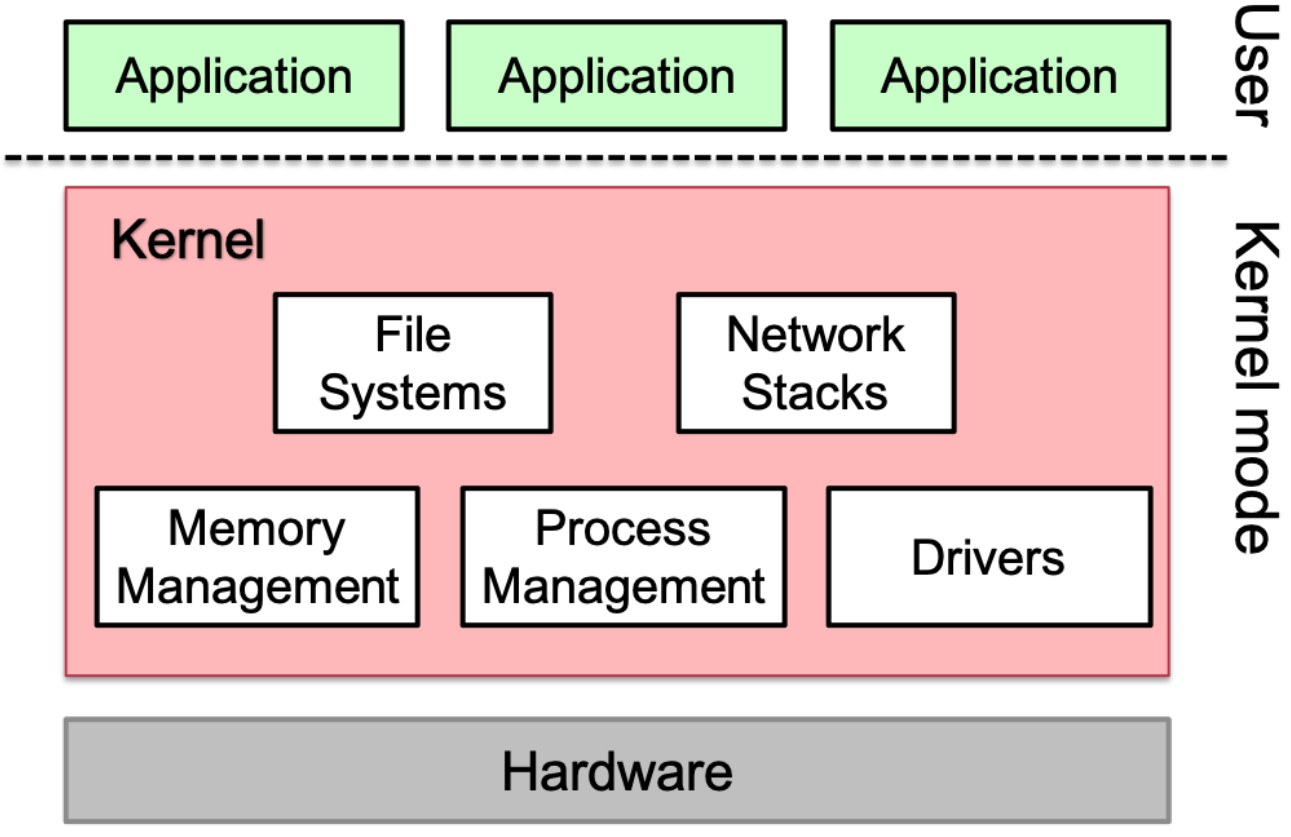
上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Harmonizing Performance and Isolation in Microkernels with Efficient Intra-kernel Isolation and Communication

Jinyu Gu, Xinyue Wu, Wentai Li, Nian Liu, Zeyu Mi,

Yubin Xia, Haibo Chen

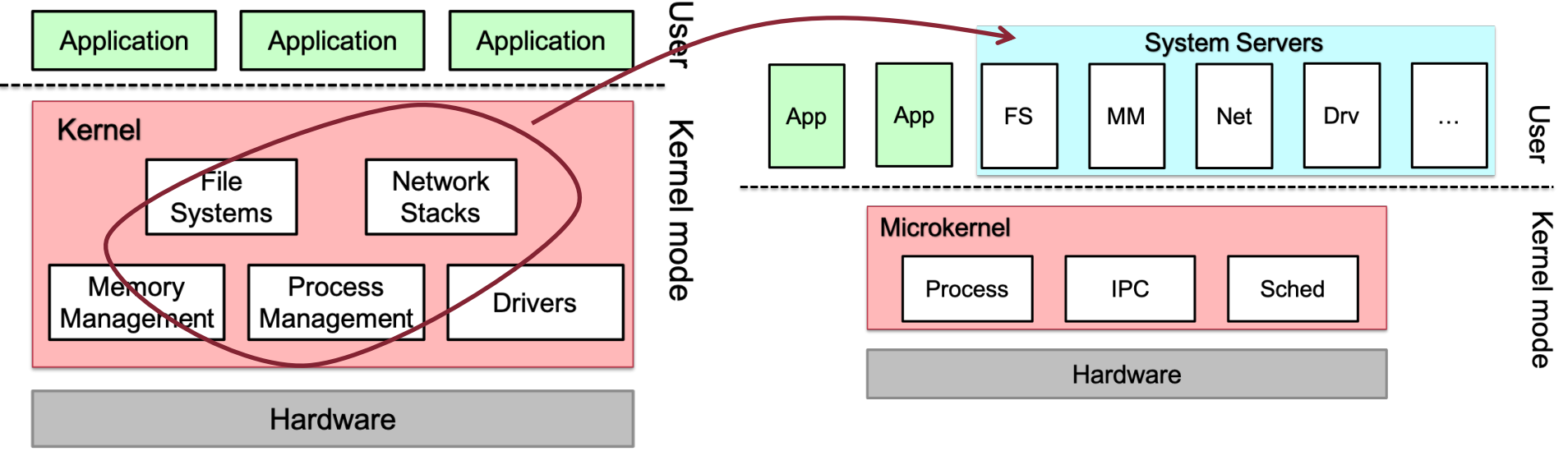
Monolithic Kernel and Microkernel



Monolithic Kernel and Microkernel

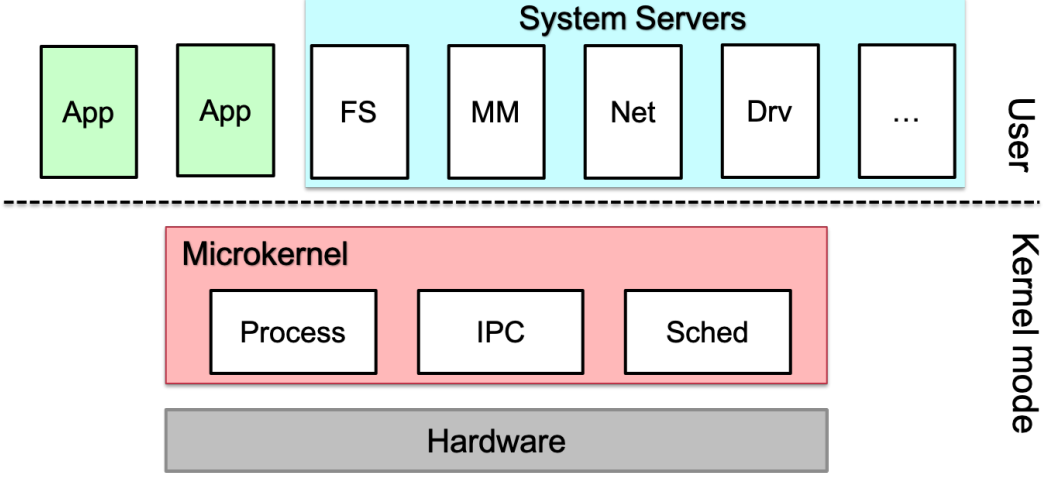
Microkernel's philosophy:

Moving most OS components into isolated user processes



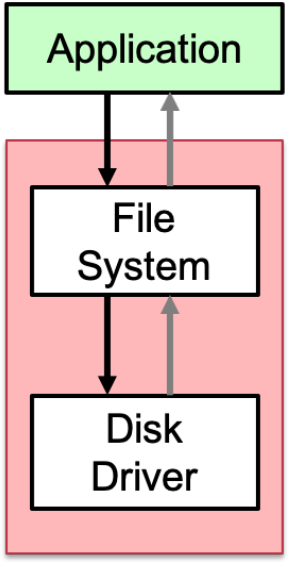
Benefits and Usages of Microkernel

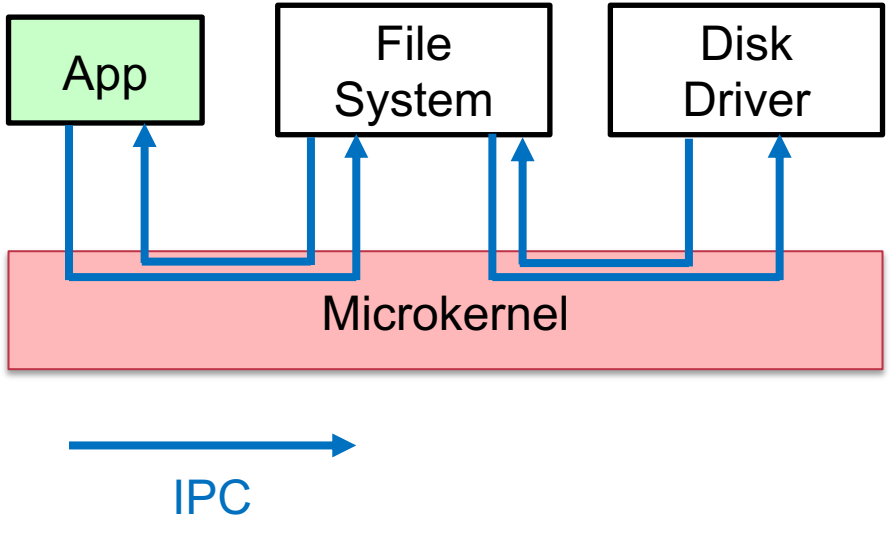
- Achieves good extensibility, security, and fault isolation
- Succeeds in safety-critical scenarios (Airplane, Car)
- For more general-purpose applications (Google Zircon)



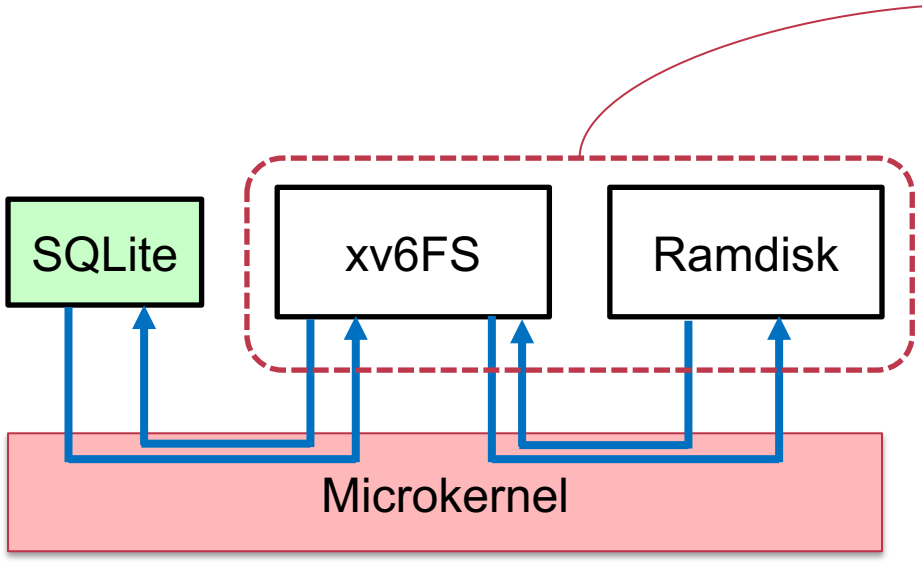
Expensive Communication Cost

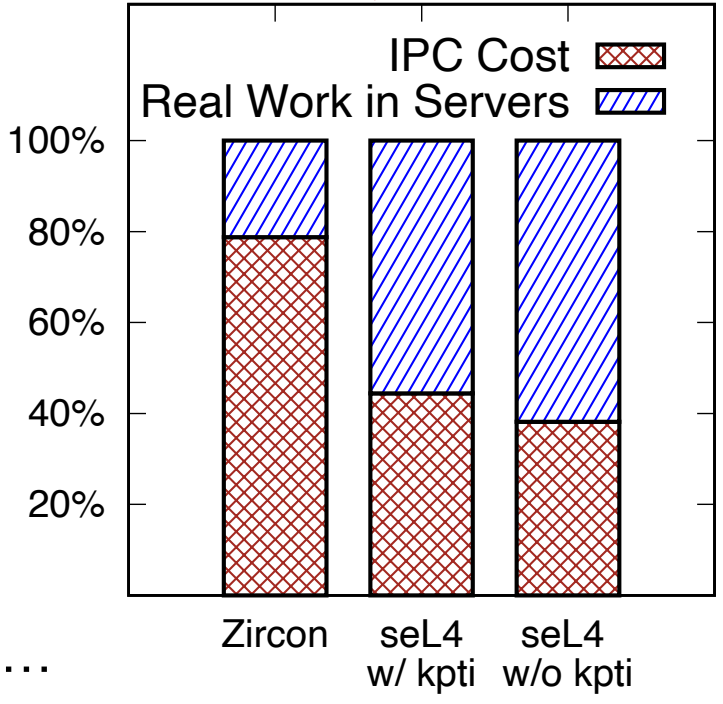
- **Tradeoff: Performance and Isolation**
  - Inter-process communication (IPC) overhead





IPC Overhead is Considerable





Direct cost: privilege switch, process switch, …

Indirect cost: CPU internal structures pollution

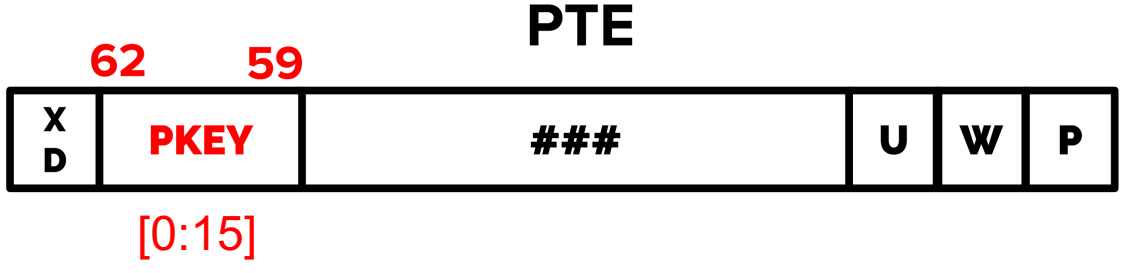
Evaluated on Dell PowerEdge R640 server with Intel Xeon Gold 6138 CPU

Goal: Both Ends

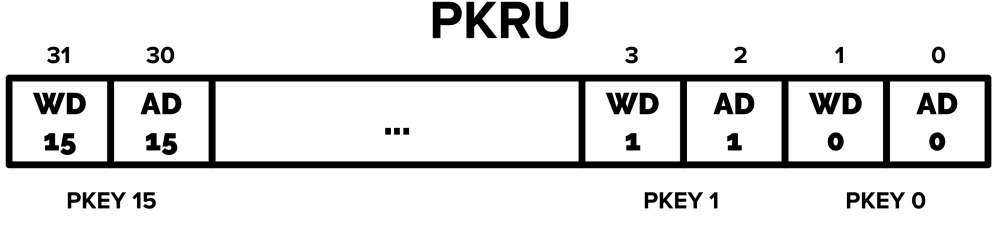
- **Harmonize the tension between Performance and Isolation in microkernels**
  - Reducing the IPC overhead
  - Maintaining the isolation guarantee

New Hardware Brings Opportunities

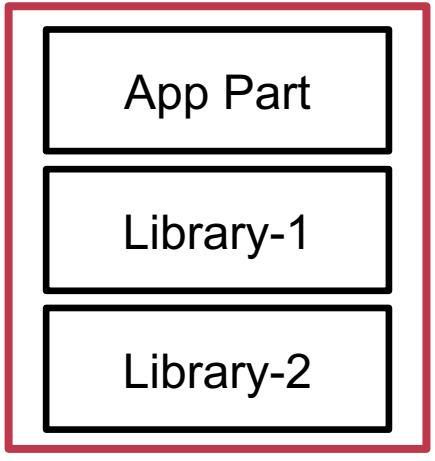
- **PKU: Protection Key for Userspace (aka. MPK)**
  - Assign each page one PKEY (i.e., memory domain ID)



  - A new register PKRU stores read/write permission

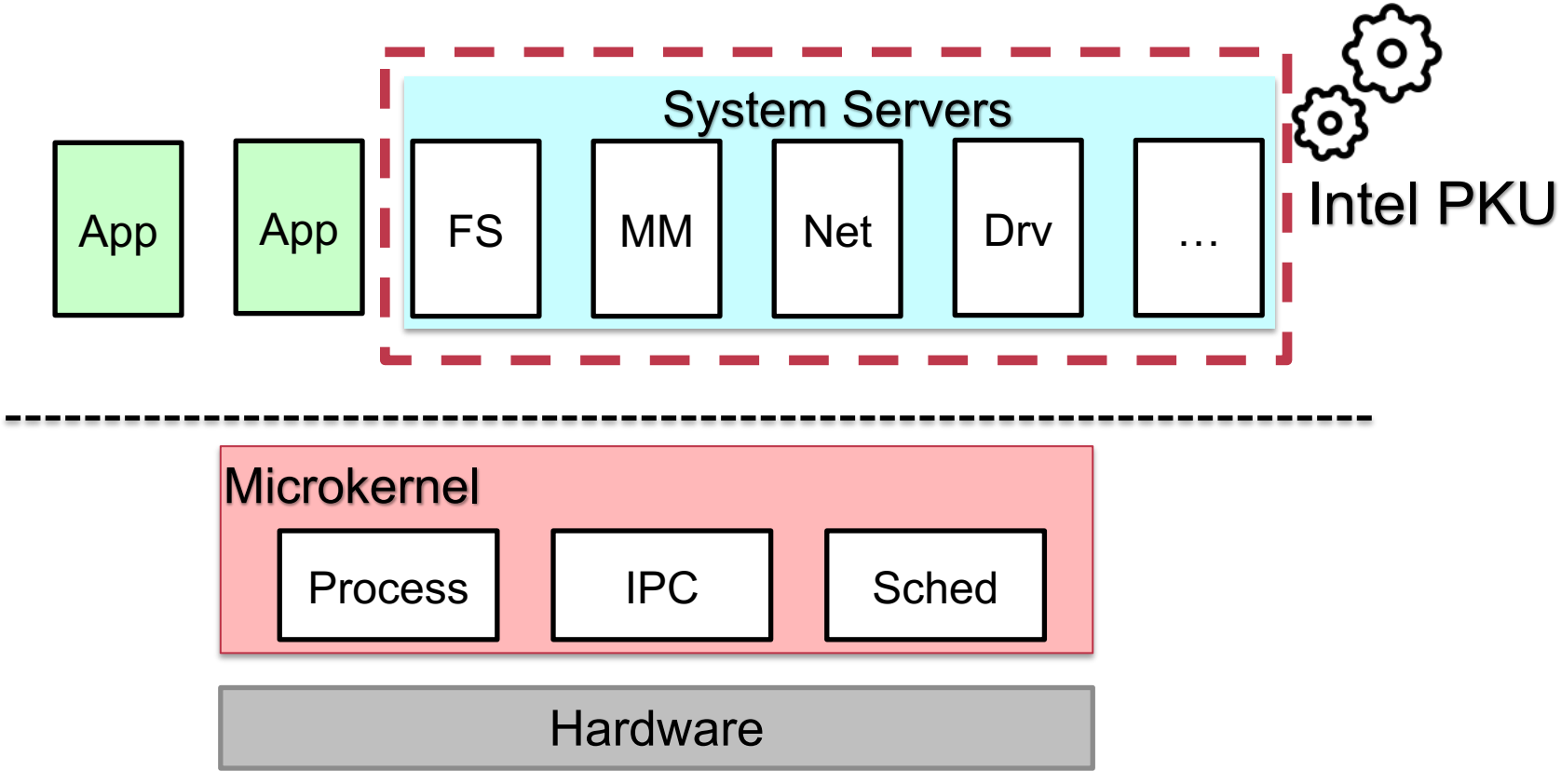


Efficient Intra-Process Isolation



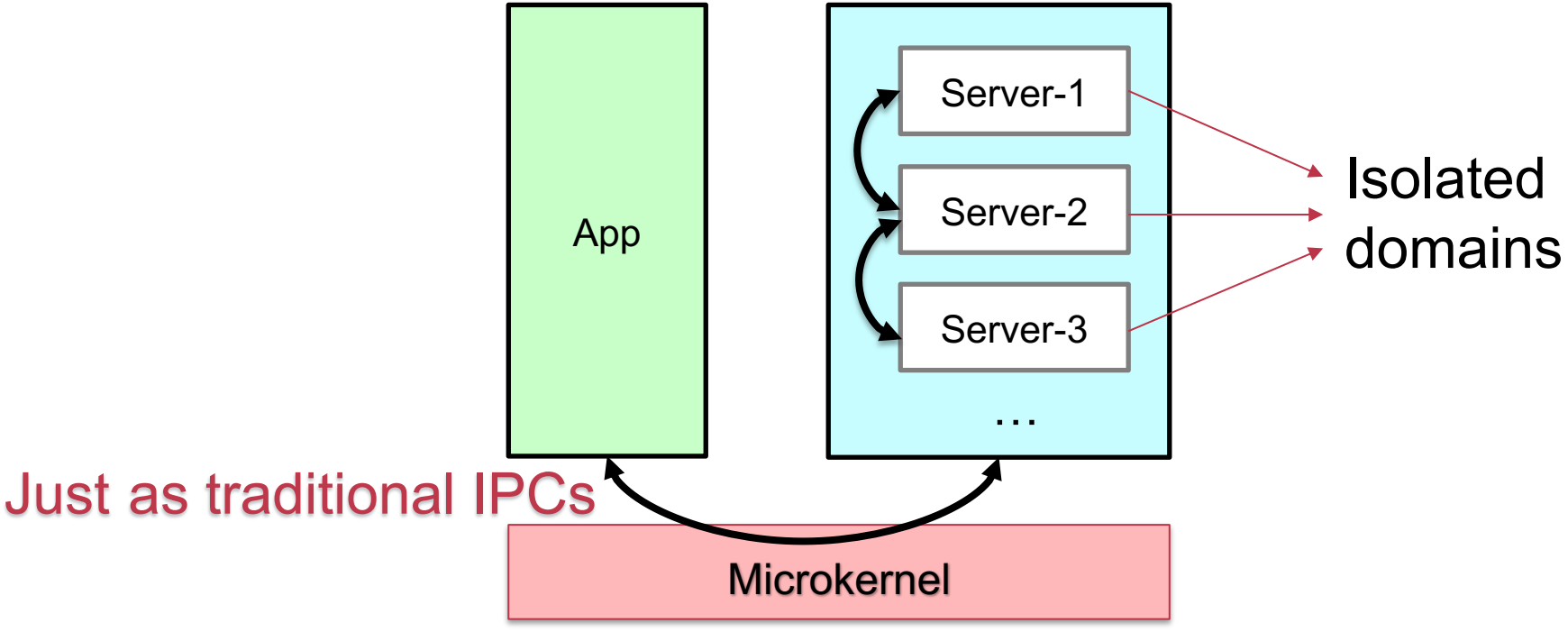
- ERIM [Security'19] & Hodor [ATC'19]
  - Based on **Intel PKU**
  - Build isolate domains in the same process **efficiently**
  - Domain switch only takes **28 cycles** (modify PKRU)

Intra-Process Isolation + Microkernel



Design Choice #1

Isolate different system servers in a single process.

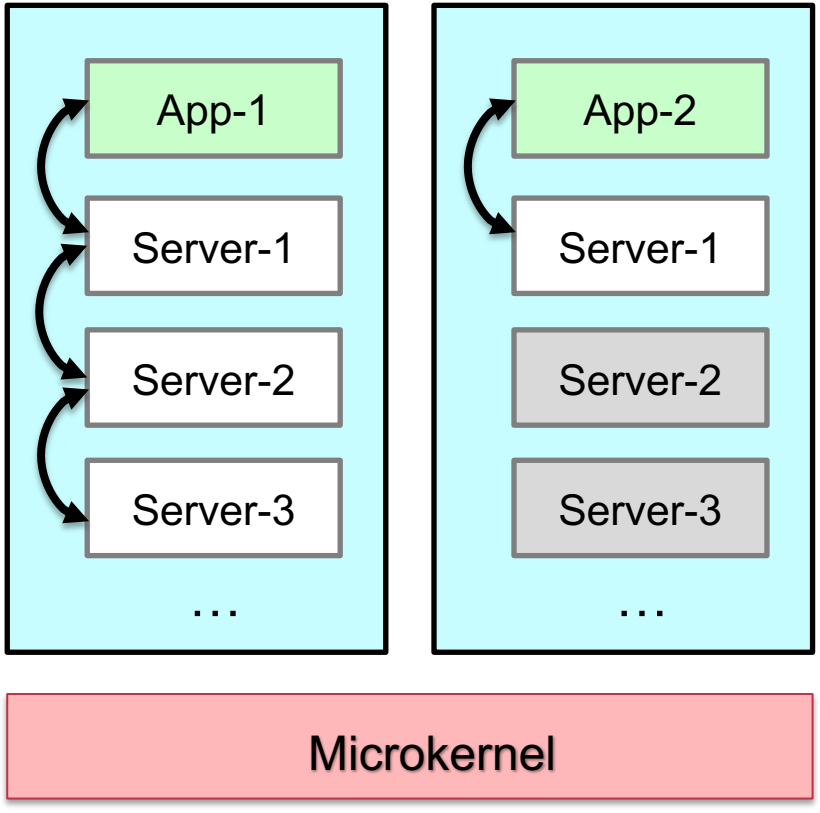


Design Choice #2

Let's get more aggressive!

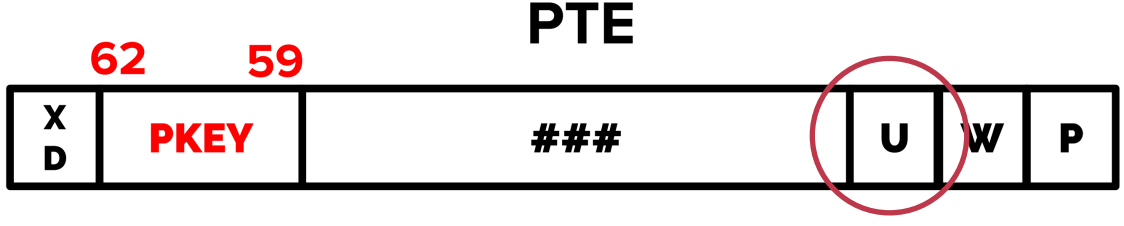
Drawbacks

1. Update Server mapping is costly
2. IPC connection is also costly
3. Less flexibility for applications on address space and using PKU

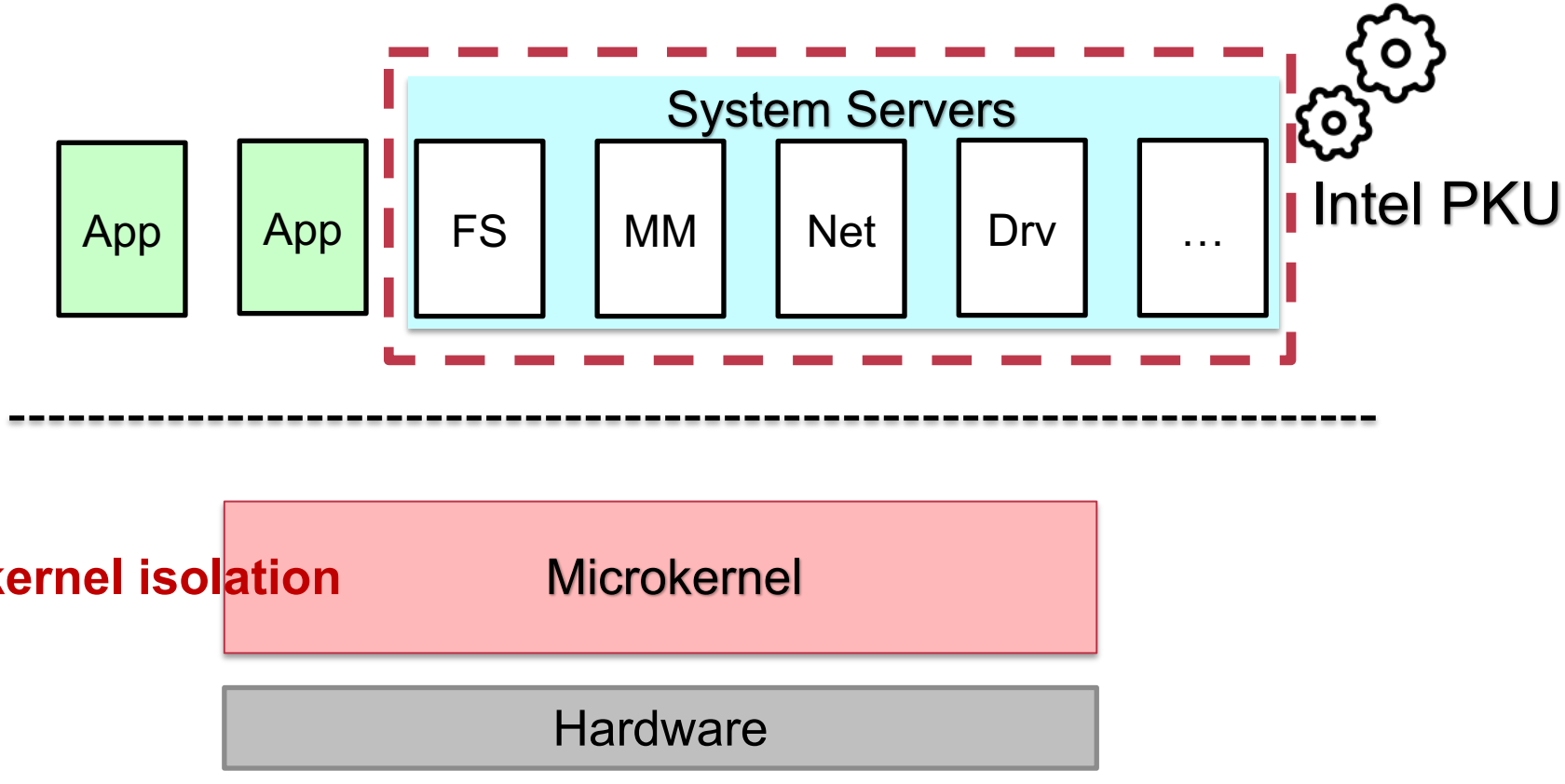


An Observation on Intel PKU

- **A misleading name**
  - Protection Key for Userspace
- **It still takes effect when in kernel (ring-0)**
  - The “Userspace” means user-accessible memory
  - U/K bit in PTE

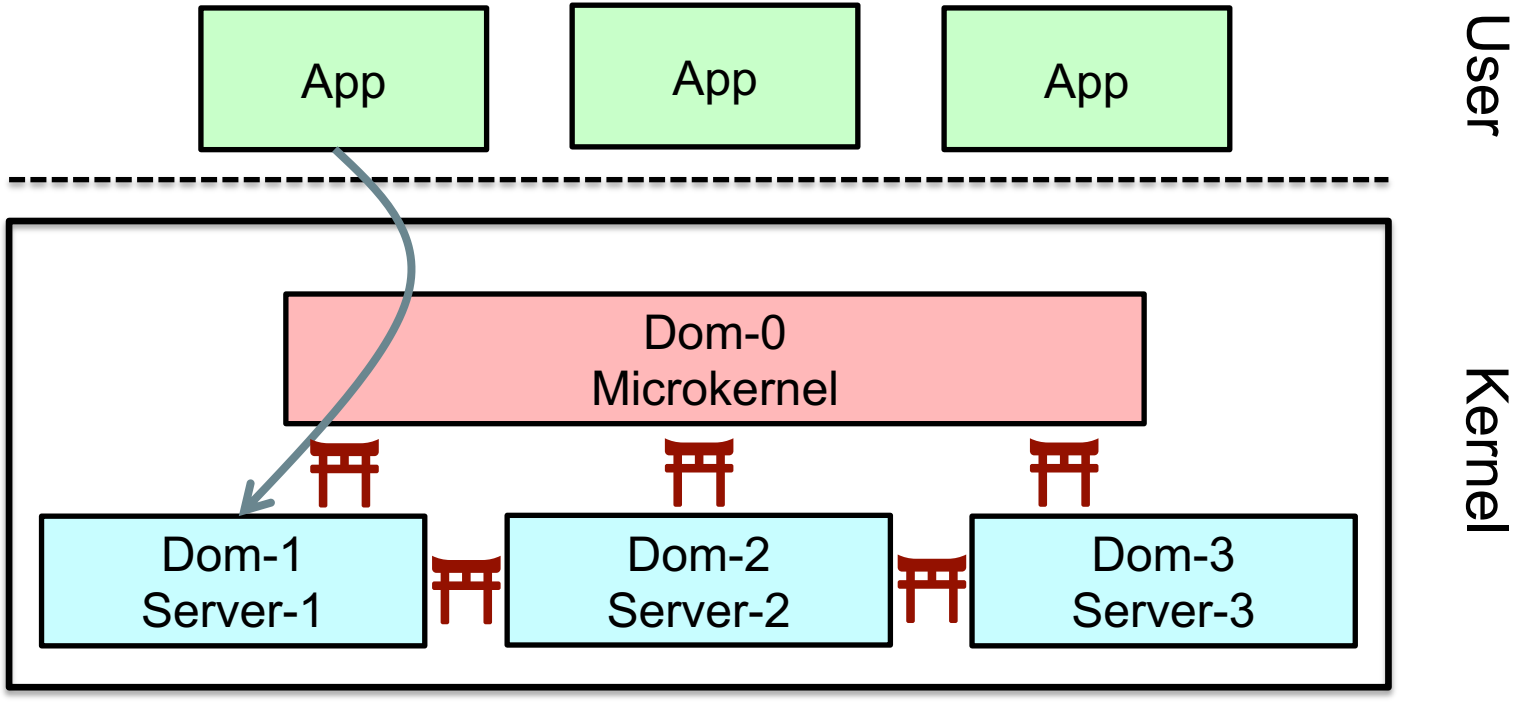


UnderBridge: Sinking System Servers



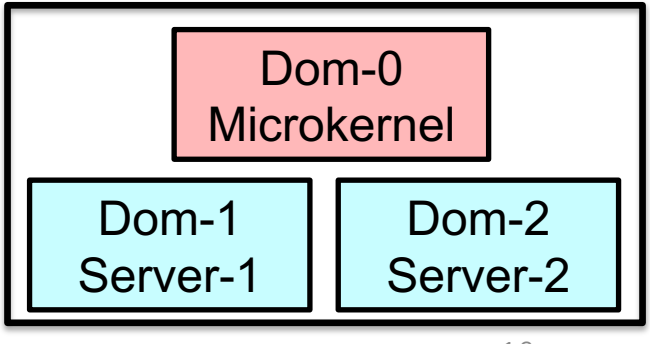
Design Choice #3: UnderBridge

- Build **execution domains** in the **kernel page table**



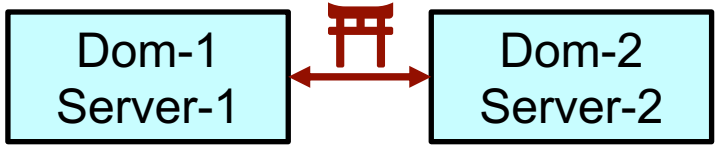
Execution Domain

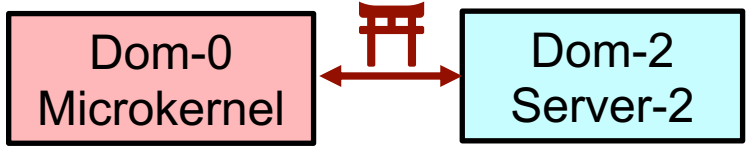
- **Execution domain 0 is for the microkernel**
  - Use memory domain 0
  - Can access all the memory
- **Others own a private memory domain**
  - A private MPK memory domain ID
- **Shared memory**
  - Allocate a free MPK memory domain ID



IPC Gate

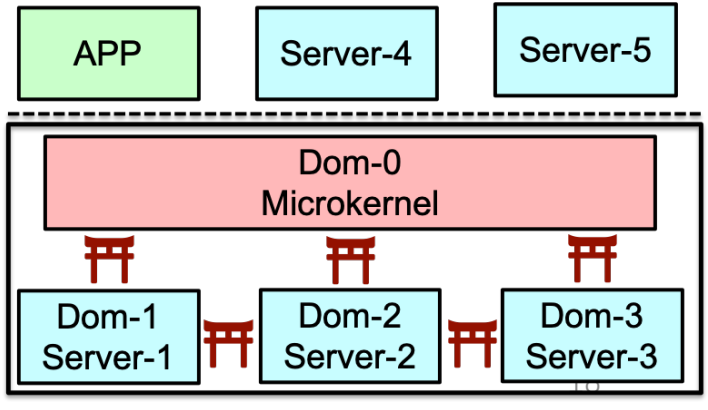
- **Connect two servers**
  - Generated by the microkernel
  - Resides in memory domain 0 (execute-only for servers)
- **Transfer control flow during IPC invocations**
  - context switch and domain switch
- **Connect the microkernel and servers**
  - System calls





Server Migration

- **The number of execution domain is limited**
  - Hardware only provides 16 memory domains
  - Time-multiplexing is expensive
- **Move servers between user and kernel space**
  - Disjoint virtual memory regions
  - Runtime migration



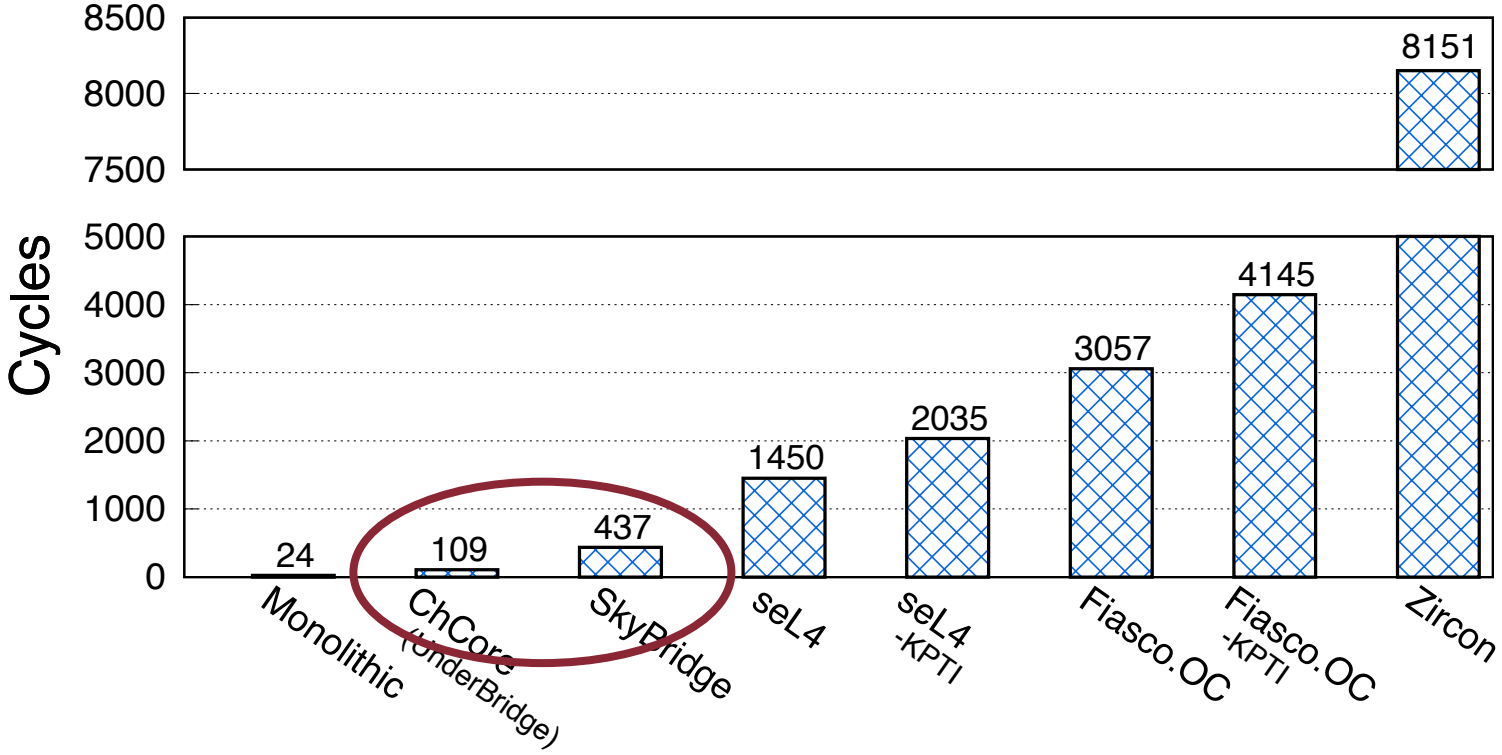
Privilege Deprivation

- **In-kernel servers have supervisor privilege**
  - Can affect the whole system if compromised
  - CFI (with binary scanning) incurs runtime overhead
  - Binary rewriting only is infeasible
- **Prevent servers to execute privilege instructions**
  - Add a tiny secure monitor in hypervisor mode
  - For instructions rarely execute: VMExits
  - For instructions that frequently required: Rewriting

Other Designs and Implementations

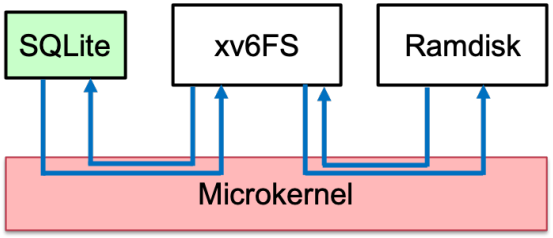
- IPC capability authentication
- Seamless server migration
- Privilege deprivation details

Cross-server IPC Round-Trip Latency

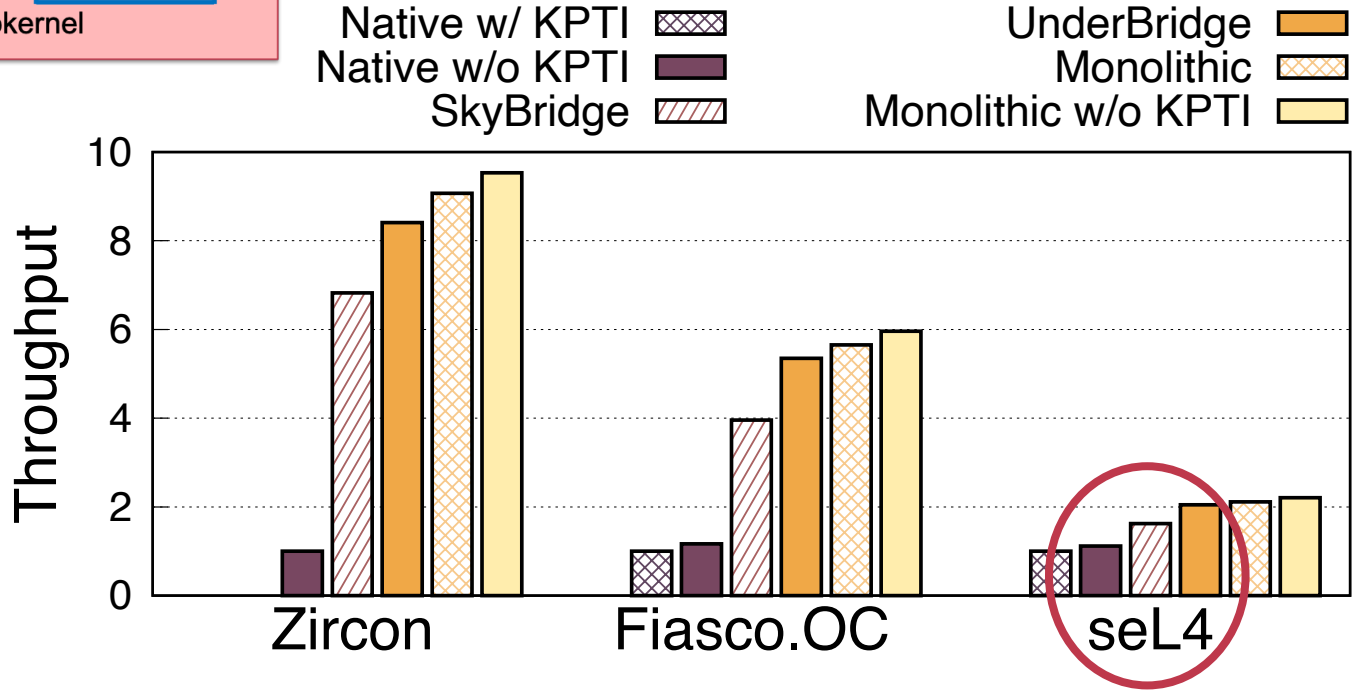


Evaluated on Dell PowerEdge R640 server with Intel Xeon Gold 6138 CPU

SQLite Throughput under YCSB-A



1×~8×



Evaluated on Dell PowerEdge R640 server with Intel Xeon Gold 6138 CPU

Conclusion & Thanks!

- **UnderBridge**
  - A redesign of the runtime structure of microkernel OSes for faster OS services
  - The efficient intra-kernel isolation mechanism may also be used to harden the isolation of monolithic kernels

Q&A: gujinyu@sjtu.edu.cn