Operating Systems and Future datacenters

Dilma M. da Silva
dilmasilva@us.ibm.com

IBM T.J. Watson Research Center
Advanced Operating Systems Group

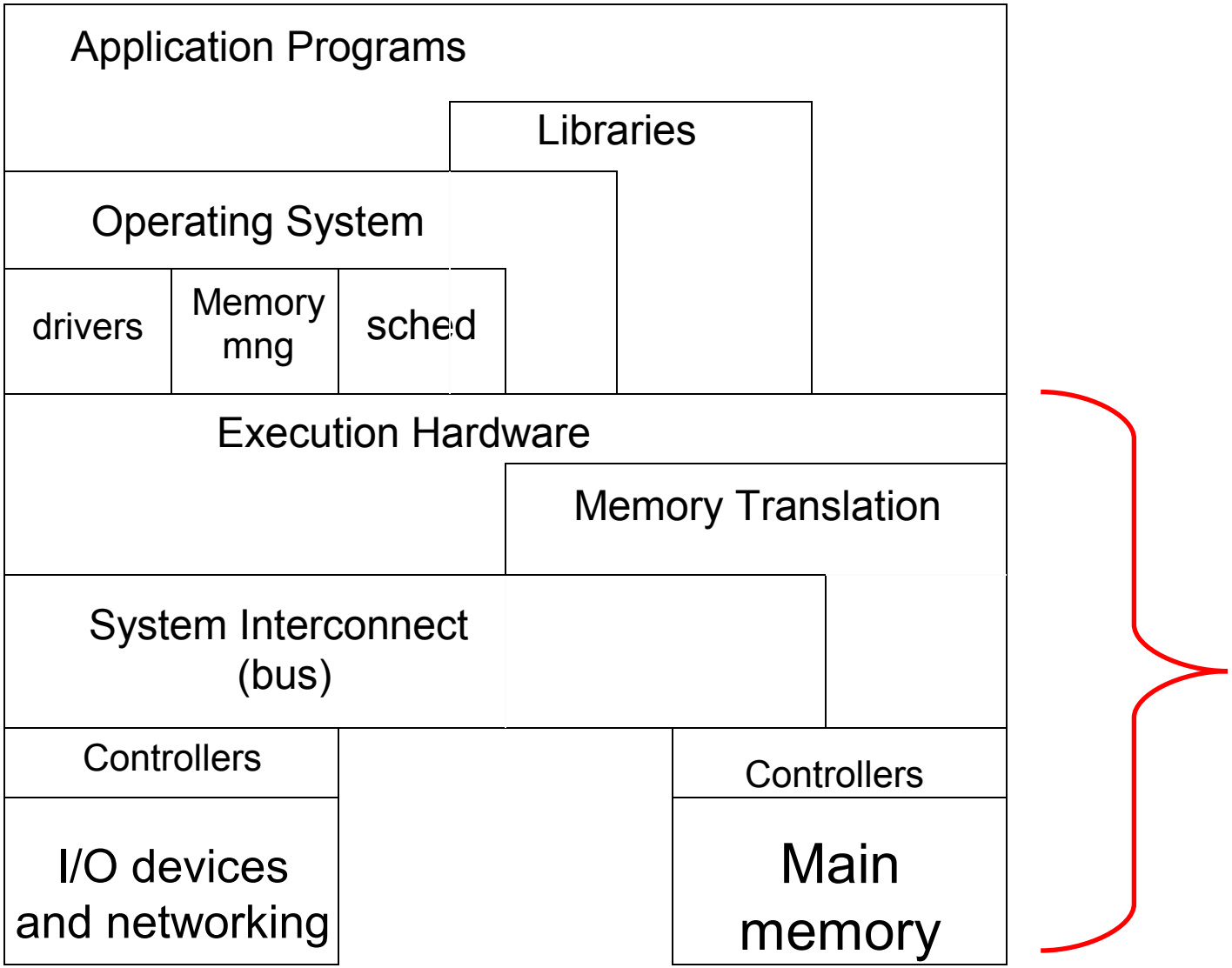
Outline – Forces of Change

- **Virtualization**
- **Total Cost of Ownership (TCO)**
  - Scale-out / Cloud Computing / Web 2.0
  - Power budget
- **Hardware evolution**
  - Multicore
  - 3D integration/packaging
- **Bug Resistance**

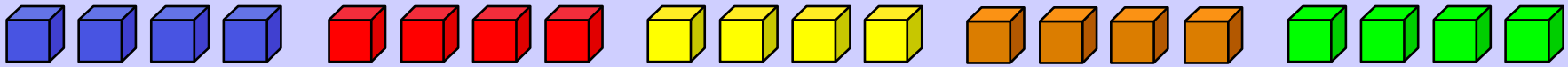
Virtualization

- **What ?**
- **Why ?**
- **Is it hard to do?**
- **Who ?**

(Recap: Architecture)



What: Virtualization Concept



Virtual Resources

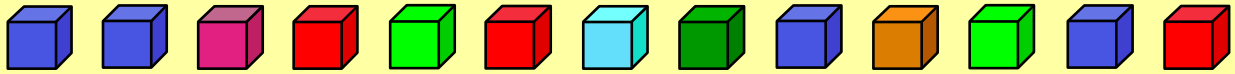
- Substitutes for real resources: same interfaces/functions, different attributes.
- Often of part of the underlying resource, but may span multiple resources.

Virtualization – a substitution process

- Creates virtual resources from real resources.
- Primarily accomplished with software and/or firmware.

Resources

- Components with architected interfaces/functions.
- Usually physical. May be centralized or distributed.
- Examples: memory, disk drives, networks, servers.

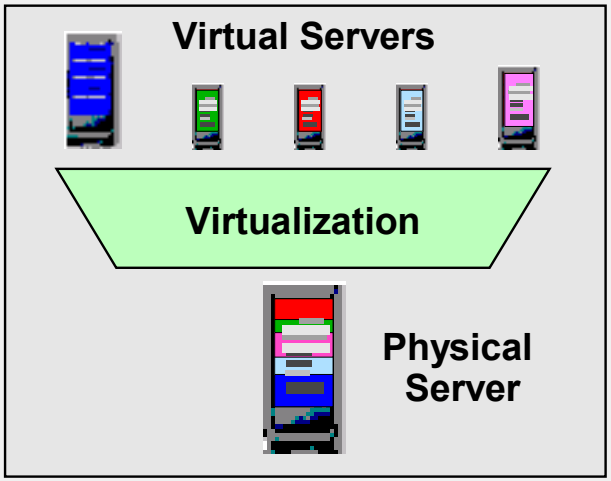


- **Separates presentation of resources to users from actual resources**
- **Aggregates pools of resources for allocation to users as virtual resources**

System Virtual Machines: why ?

- **Reduce total cost of ownership (TCO)**
  - Increased systems utilization (current servers have less than 10% average utilization, less than 50% peak utilization)
  - Reduce hardware (25% of the TCO)
  - Space, electricity, cooling (50% of the operating cost of a data center)

Roles:
- Consolidations
- Dynamic provisioning / hosting
- Workload management
- Workload isolation
- Software release migration
- Mixed production and test
- Mixed OS types/releases
- Reconfigurable clusters
- Low-cost backup servers



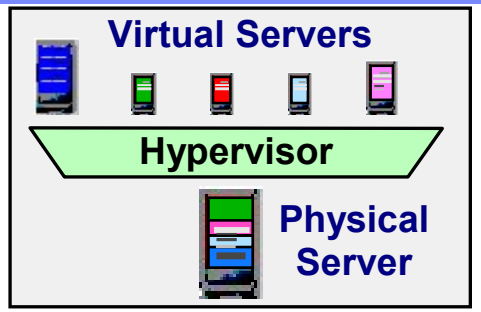
Benefits:
- Higher resource utilization
- Greater usage flexibility
- Improved workload QoS
- Higher availability / security
- Lower cost of availability
- Lower management costs
- Improved interoperability
- Legacy compatibility
- Investment protection

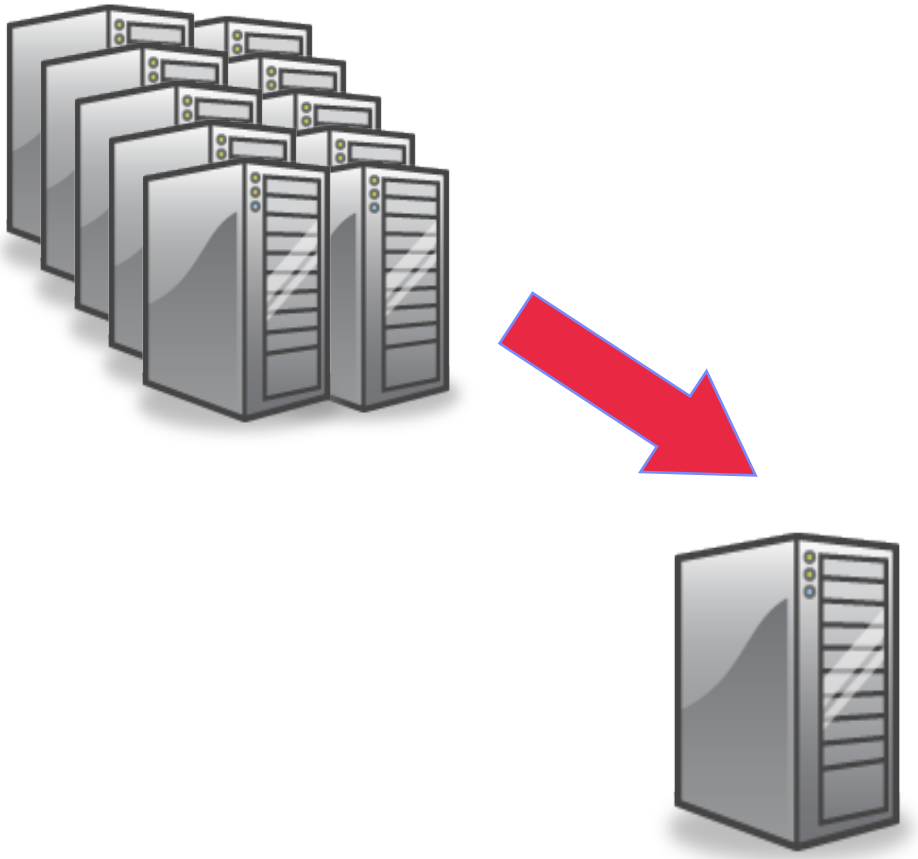
In the final analysis, the virtualization benefits take three forms:

- **Reduced hardware costs**
  - Higher physical resource utilization
  - Smaller footprints
- **Improved flexibility and responsiveness**
  - Virtual resources can be adjusted dynamically to m and to optimize service level achievement
  - Virtualization is a key enabler of on demand operat
- **Reduced management costs**
  - Fewer physical servers to manage
  - Many common management tasks become much easier

- However, server virtualization introduces some complexity and requires skills
- This partially offsets the benefits, but the net gains are generally substantial

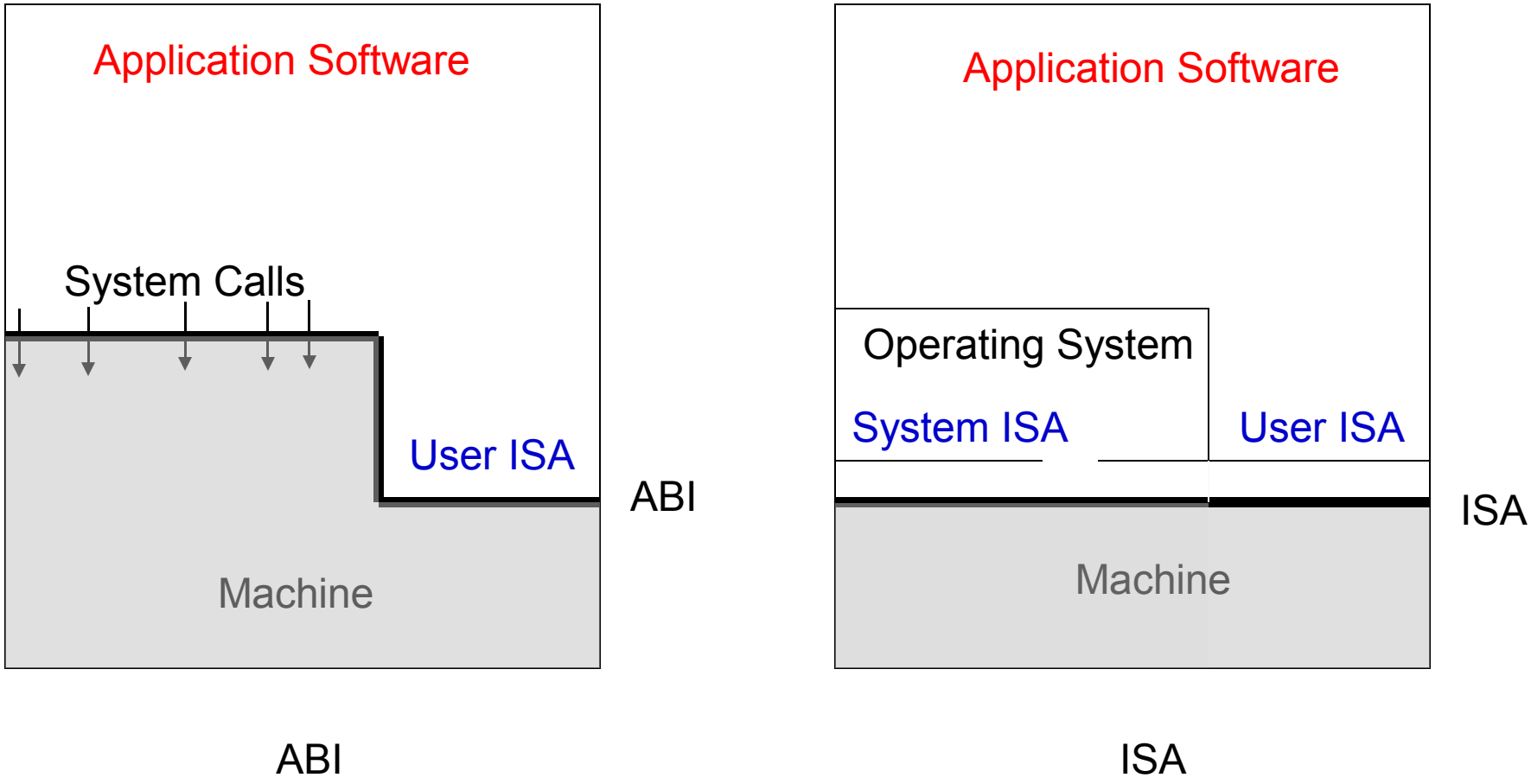
Why: Data Center Consolidation





- Security
- Isolation
- EE independence
  - Legacy apps
- Old !

Is this hard to do? (Recap: Machine Interfaces)



How we got here: VMware trajetory

- **Academic heritage: Stanford's Disco, Cellular Disco**
- **Processor virtualization:**
  - IA-32 has 17 instructions that are critical
  - VMMonitor scans instruction stream and detects the presence of instructions such as popfd
  - The instruction is replaced with code that takes the processor into privileged state and emulates the action of original code
- **Early-on success**
- **Facilities to checkpoint/restart/migrate/manage virtual machines**
- **(Acquired by EMC; recent IPO)**

How we got here: Open Hypervisors

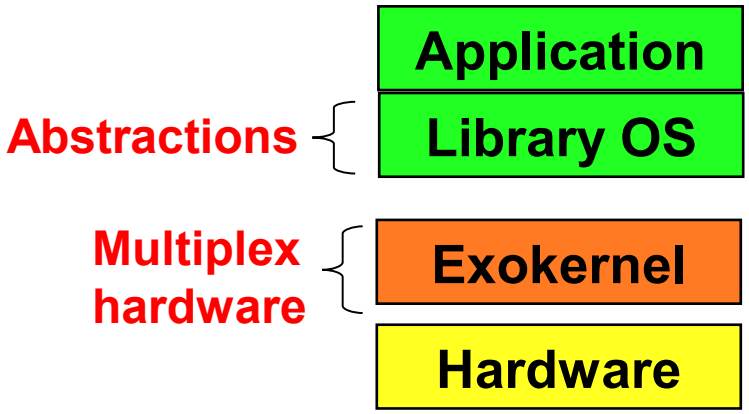
- **Xen**
  - OSDI'03
  - Paravirtualization:
    - Modify guest OS to cooperate with VMM
    - Running windows images a challenge
  - Open-source
  - Didn't get changes into Linux as expected
- **(Hardware-assist: Intel VT-x, AMD SVM)**
- **KVM**
  - Linux as the hypervisor on virtualization-enabled hardware
  - Changes got into the linux kernel quickly
  - Still has to catch up with Xen e.g. migration support; I/O virtualization
- **Which one do you embrace?**
- **Recent announcements by Oracle and Sun may play a role**
- **Microsoft Veridian**

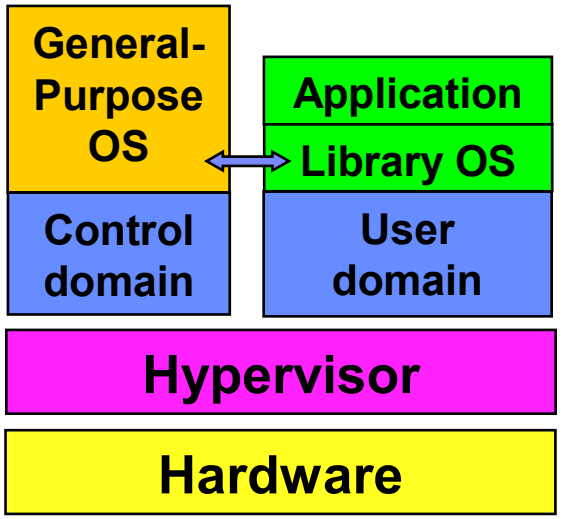
OS Research & Virtualization

- **I/O optimization**
- **Explore extra level of indirection**
  - Isolation of devices (SOSP'03, OSDI'04)
  - Time travel for debugging or intrusion detection (OSDI'04, SOSP'05)
  - "Bugs as allergies" (SOSP'05)
  - Security
  - Improving availability
  - Freedom from POSIX!!!! (HotOS'07)

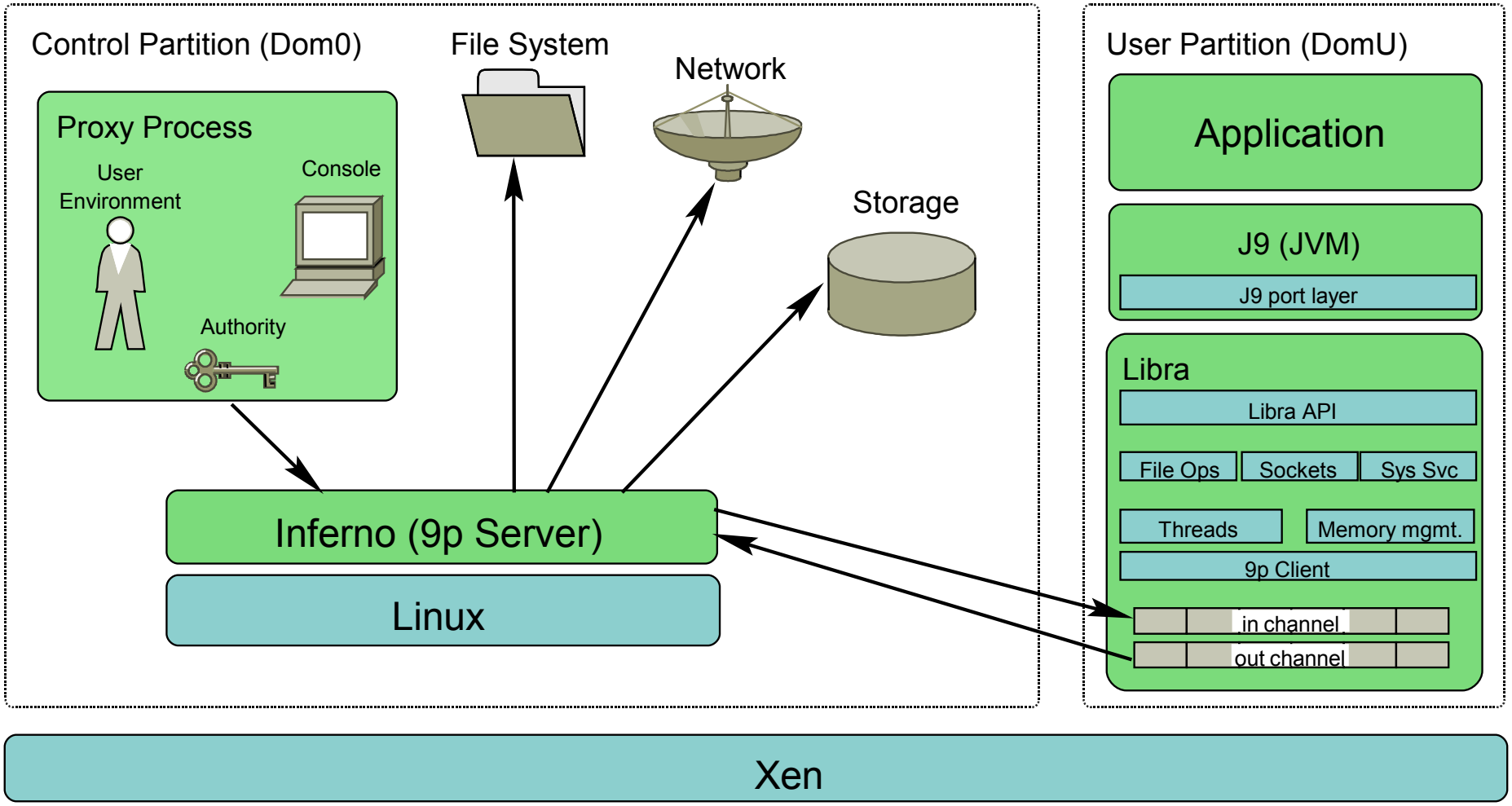
Ours: Specialized Execution Environments

- **Customized operating system support for applications**
- **Previous approaches**
  - SPIN, Vino, Scout, K42
  - Exokernel
- **Virtualization – new opportunity**

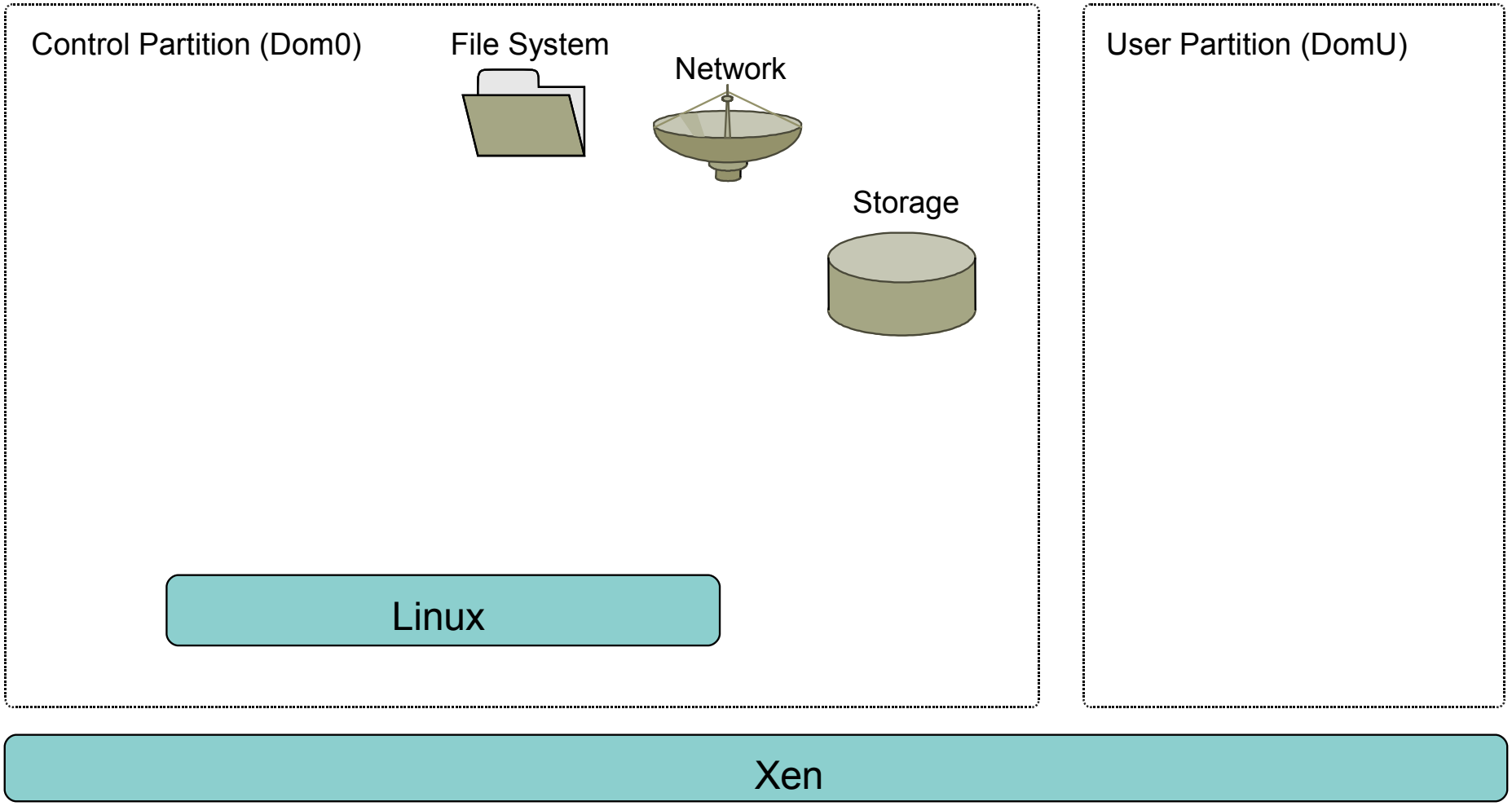




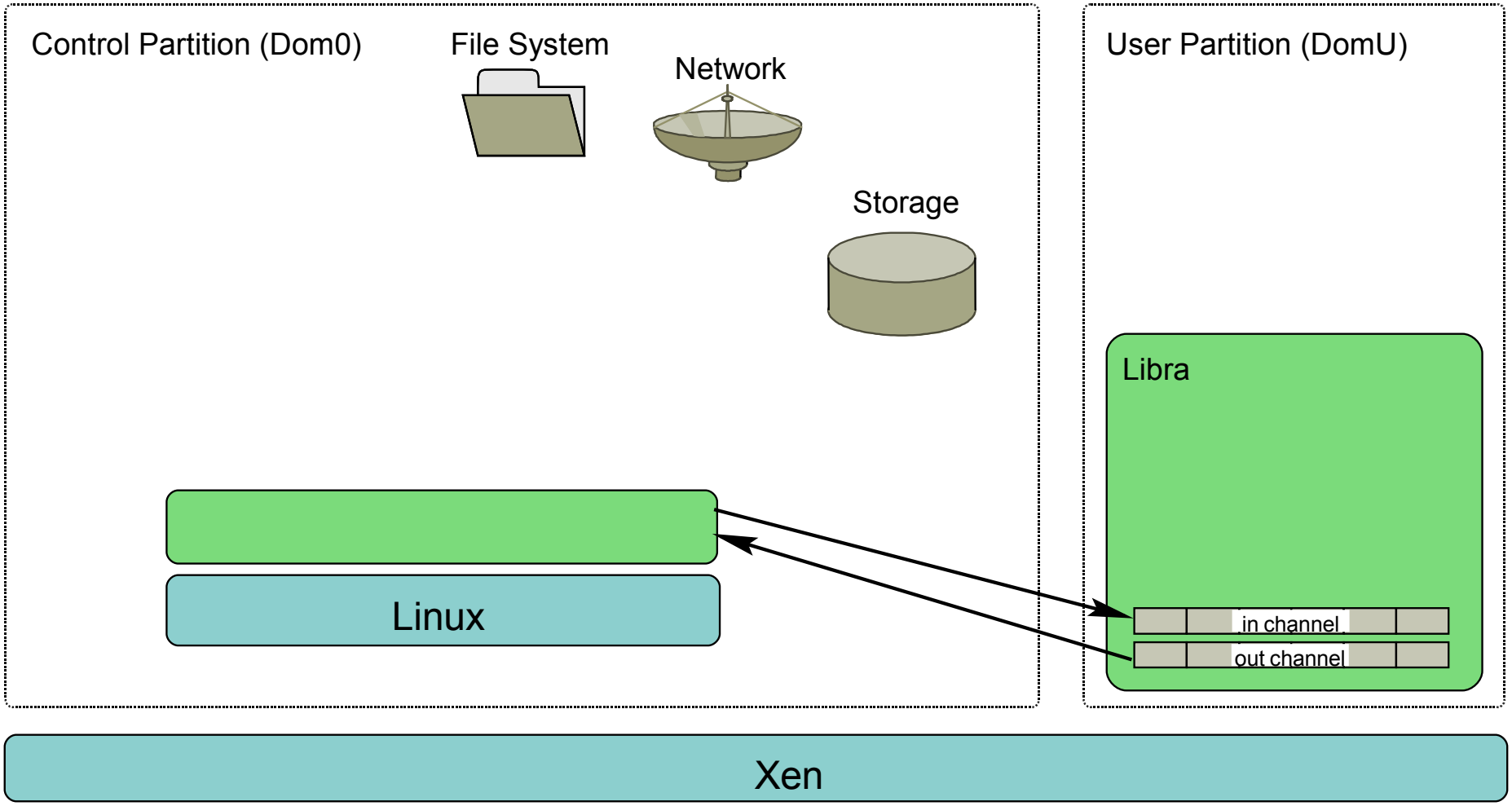
J9/Libra Architecture



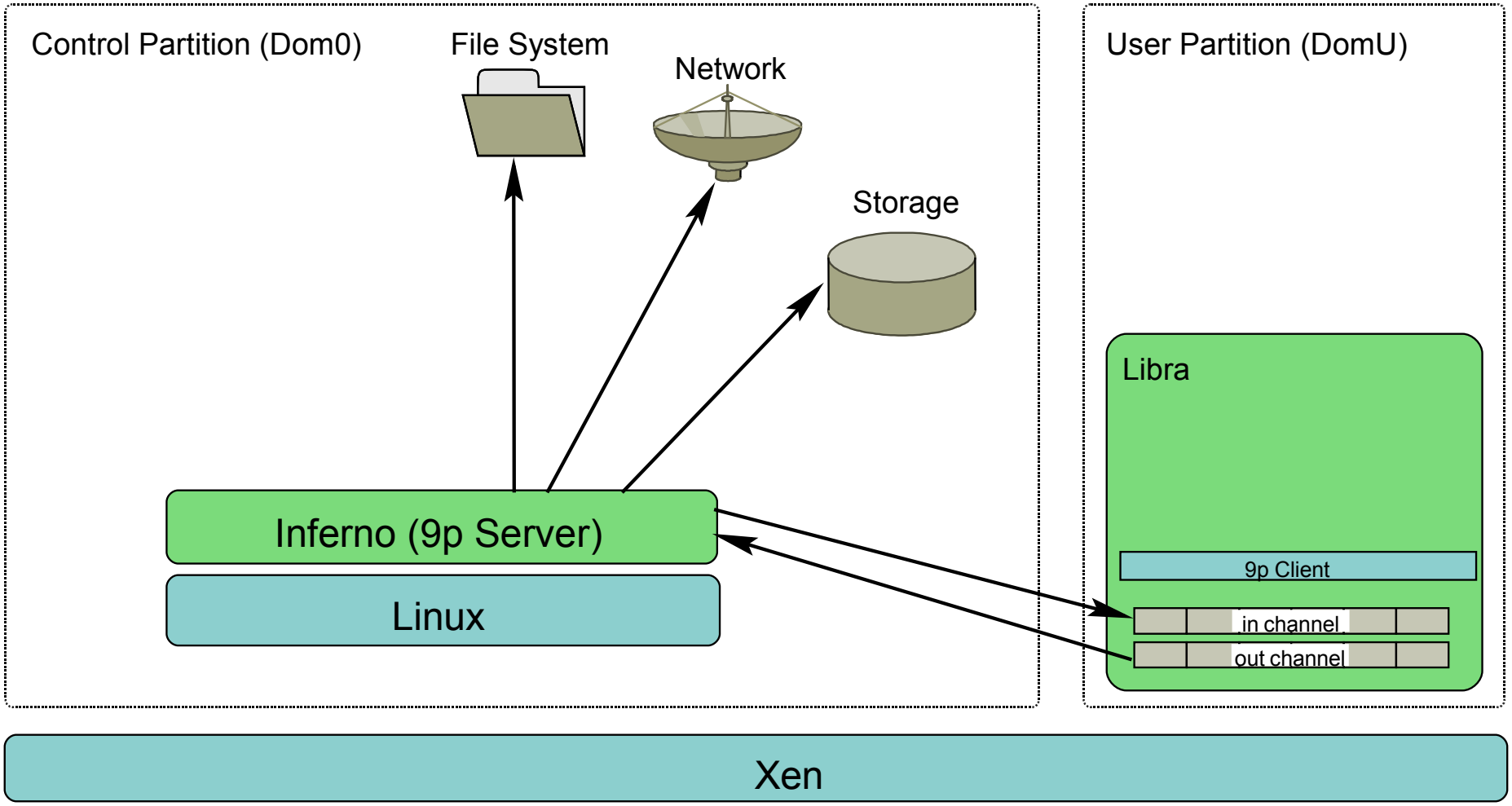
J9/Libra Architecture



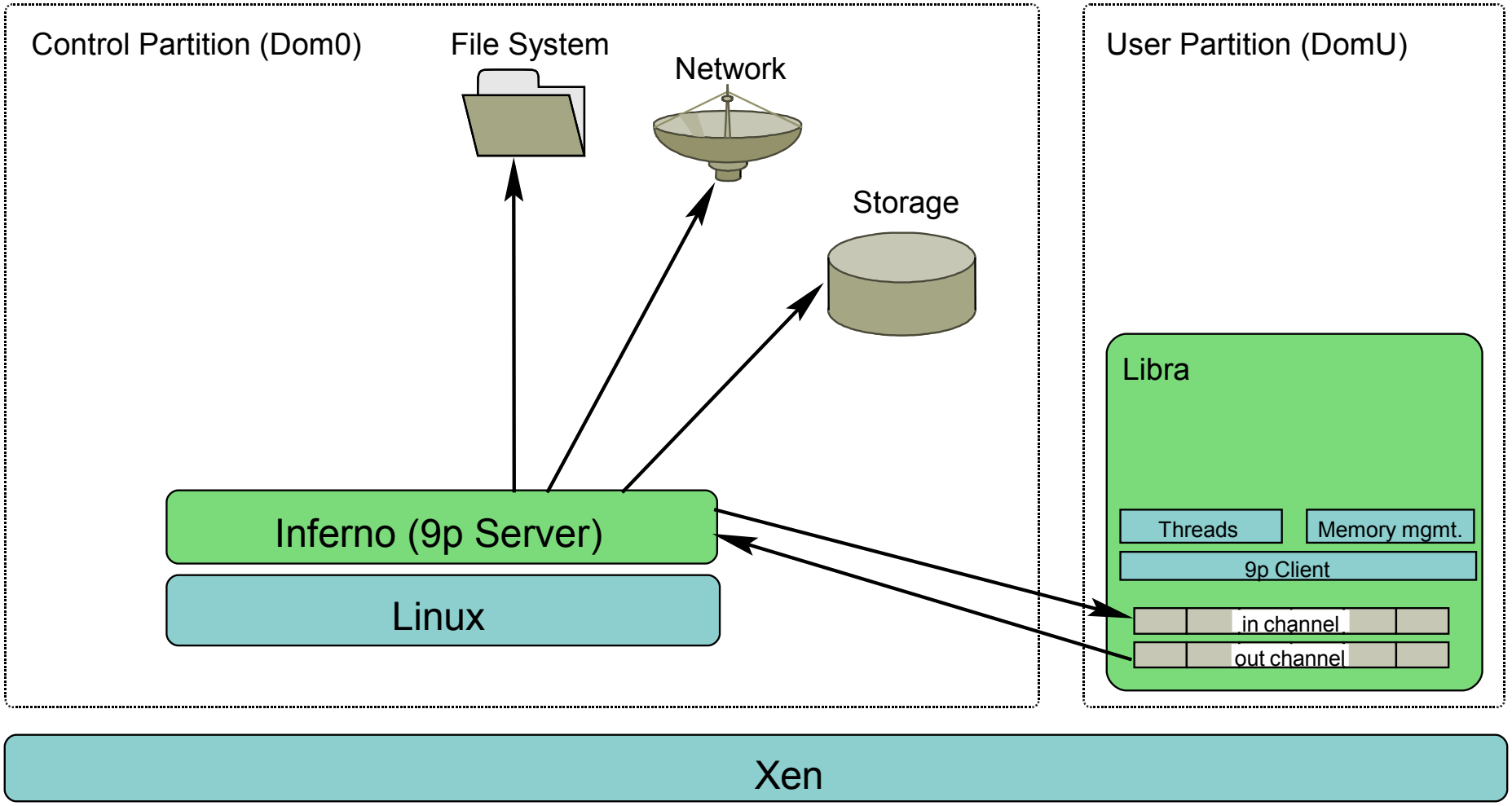
J9/Libra Architecture



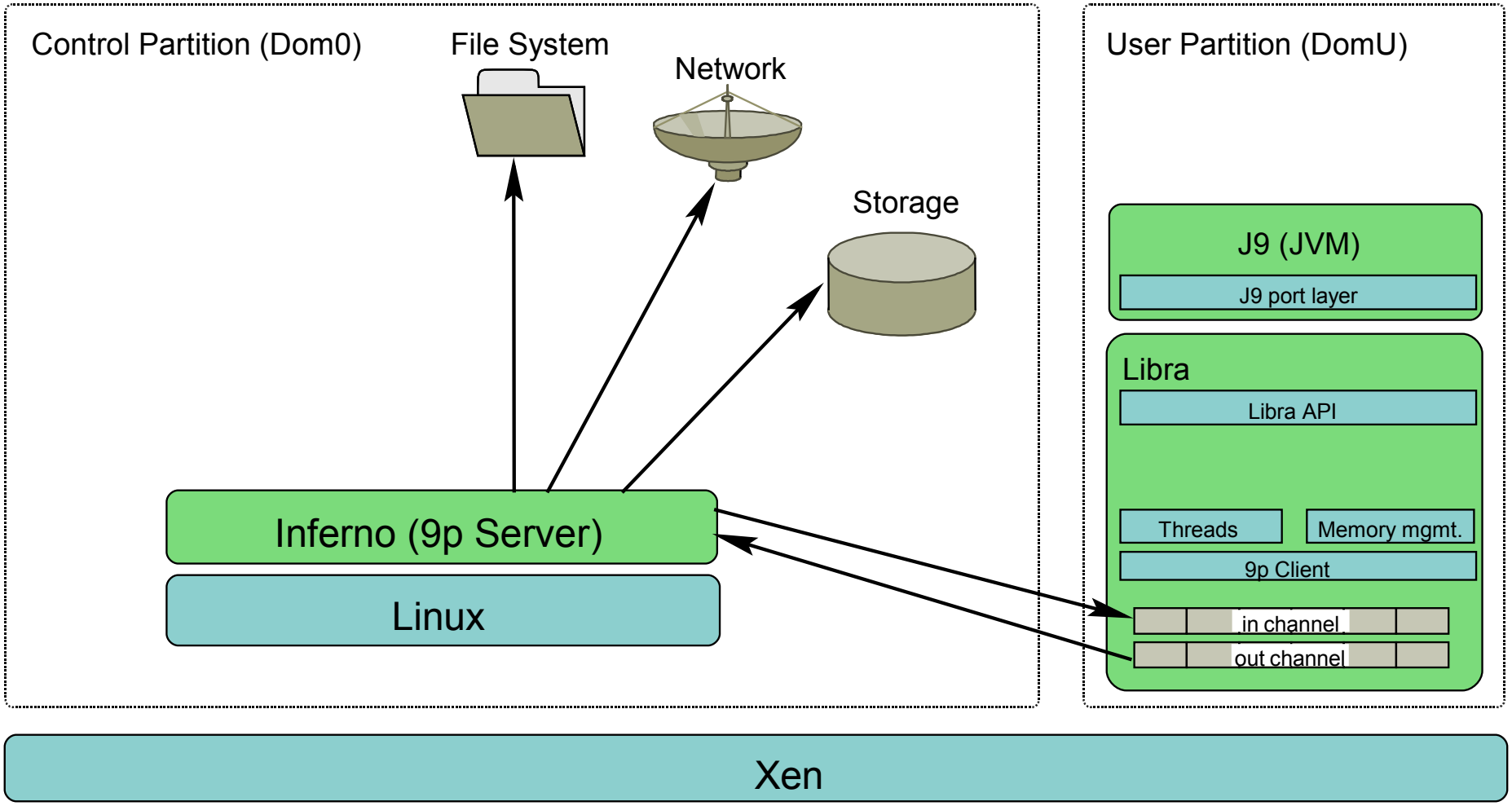
J9/Libra Architecture



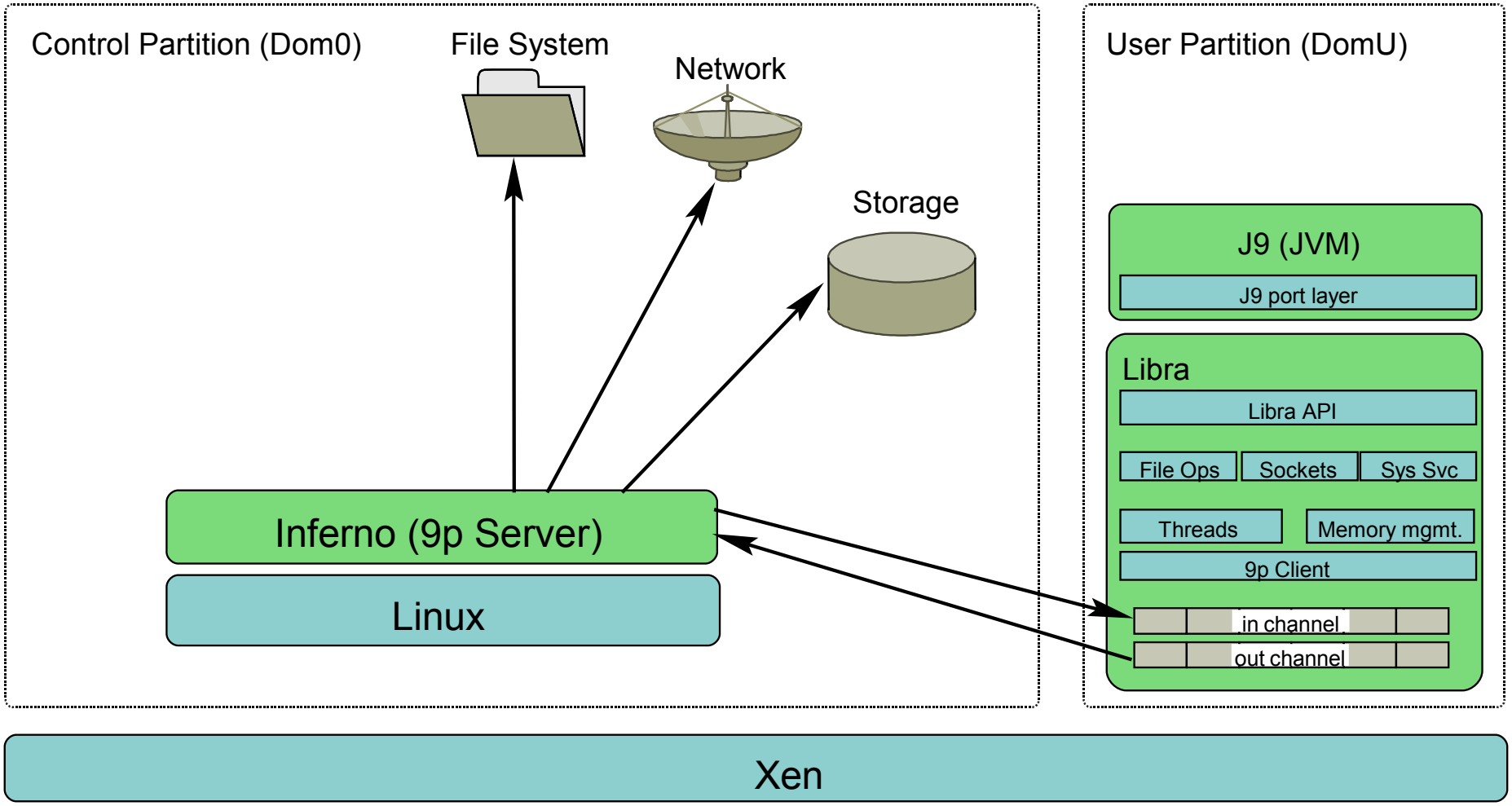
J9/Libra Architecture



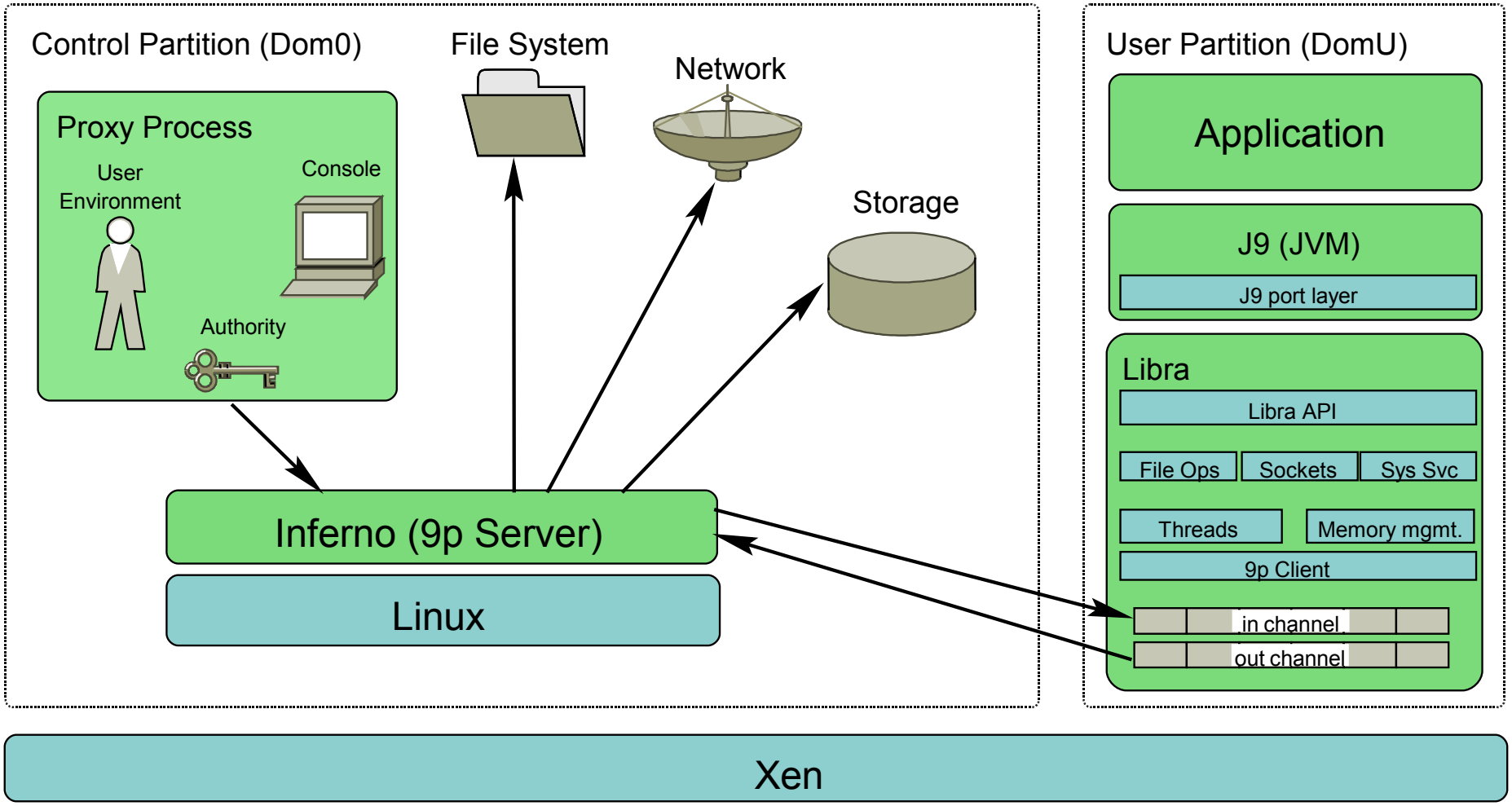
J9/Libra Architecture



J9/Libra Architecture



J9/Libra Architecture



Libra optimizations: file caching

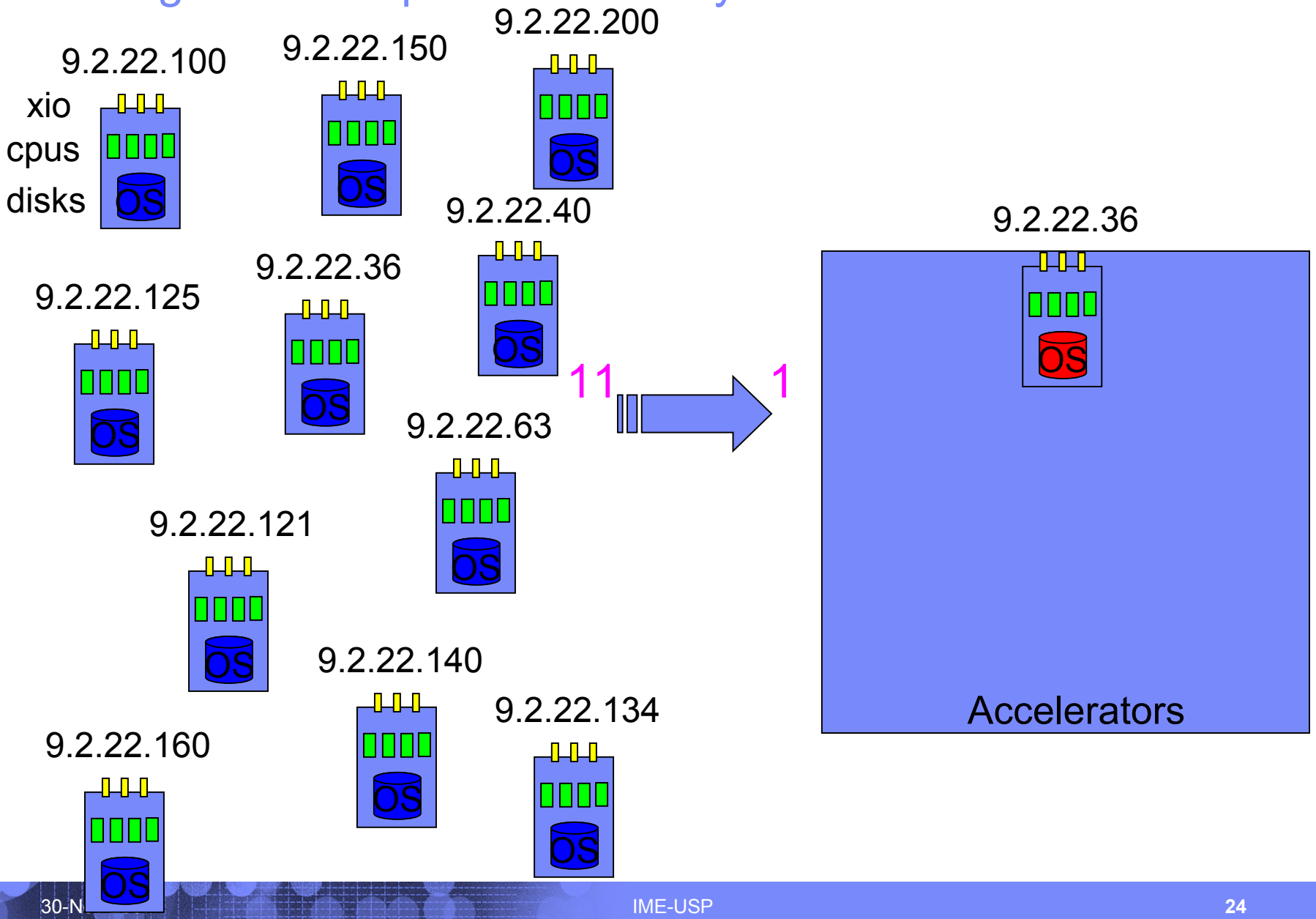
- **Index and some raw data must be in memory**
- **Nutch query back-end relies on OS buffer cache**
- **Going to control partition is expensive**
- **Solution: cache files locally in Libra**
- **Average lseek() & read() cost for back-end:**
  - J9/Linux: 2.25 usec
  - J9/Libra: 0.9 usec

Libra optimizations: socket streaming

- **Nutch query back-end is a streaming application**
- **Requests buffered in control partition**
- **Fetching them on-demand adds latency**
- **Sending results ties up worker threads**
- **Solution: stage socket data into/out of Libra partition**
  - New requests are always available locally
  - Results are sent asynchronously in batches

- **(Performance data in ACM VEE'07)**
- **Currently we support C/C++/glibc on both ppc and x86-64**

Management simplification: Many Become One



```
$ ssh chassis0
chassis0 > java HelloWorld
```

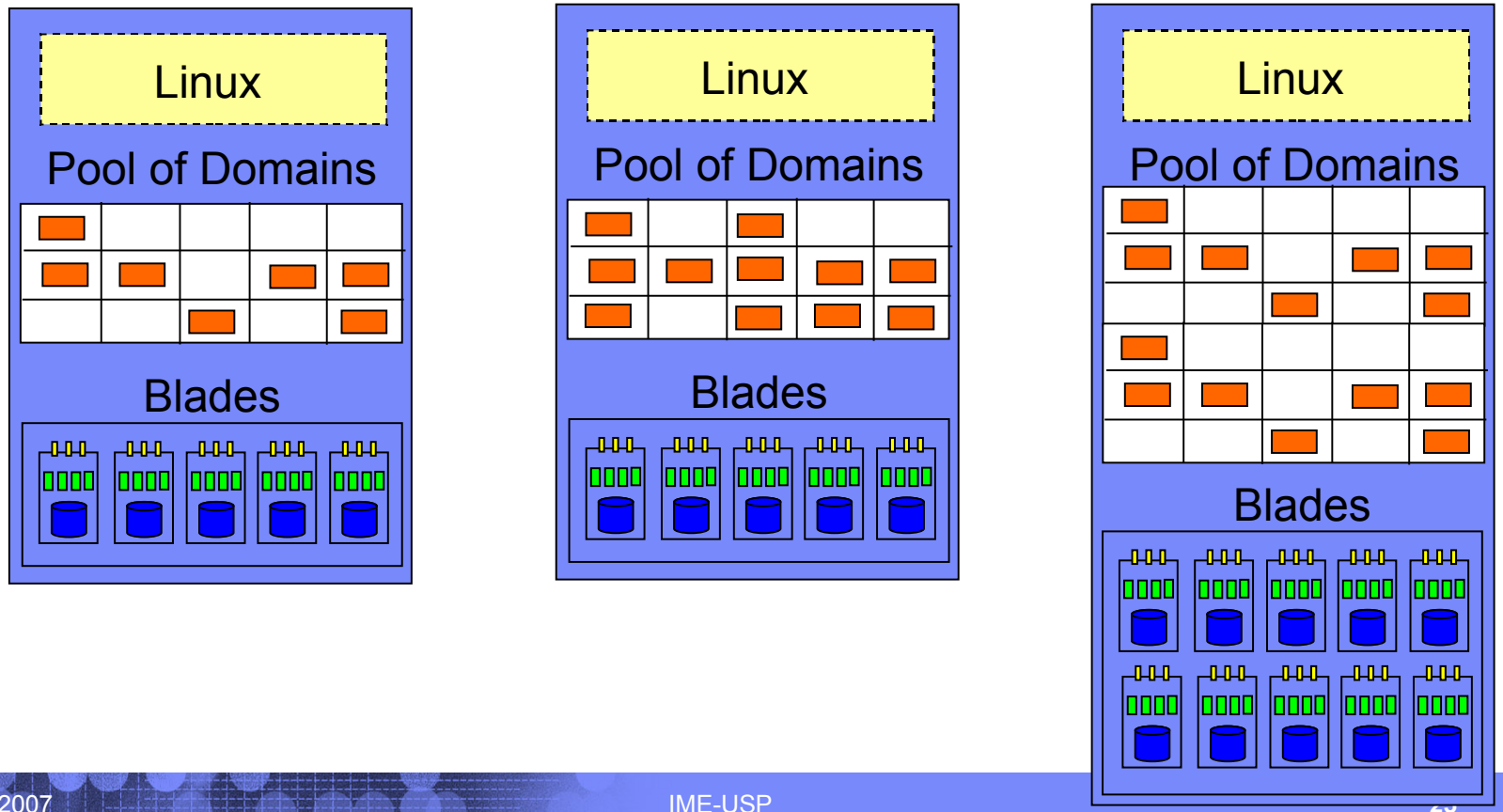
```
$ ssh chassis1
chassis1 > java HelloWorld
```

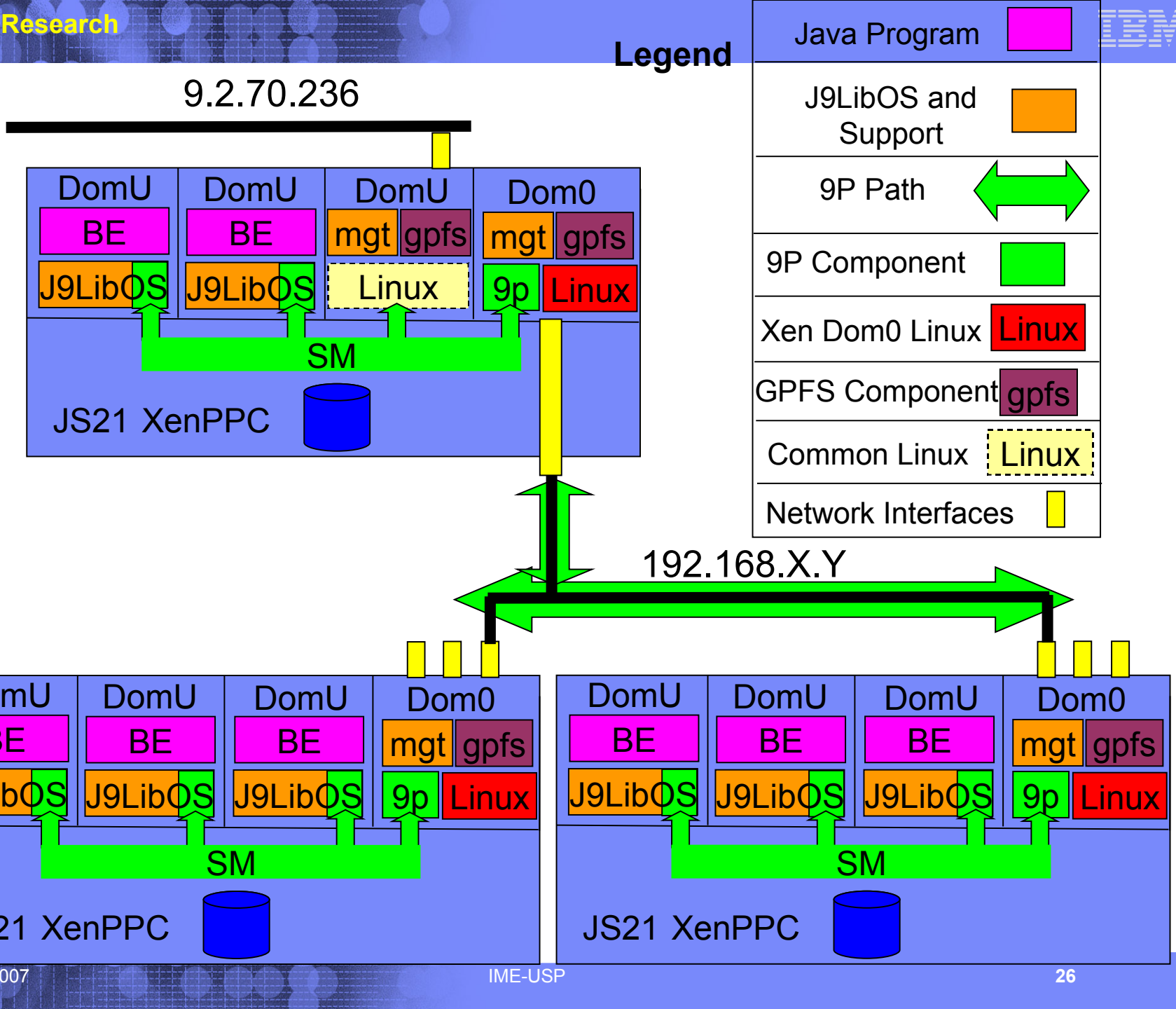
```
$ ssh chassis2
chassis2 > java HelloWorld
```

Virtual Chassis 0

Virtual Chassis 1

Virtual Chassis 2

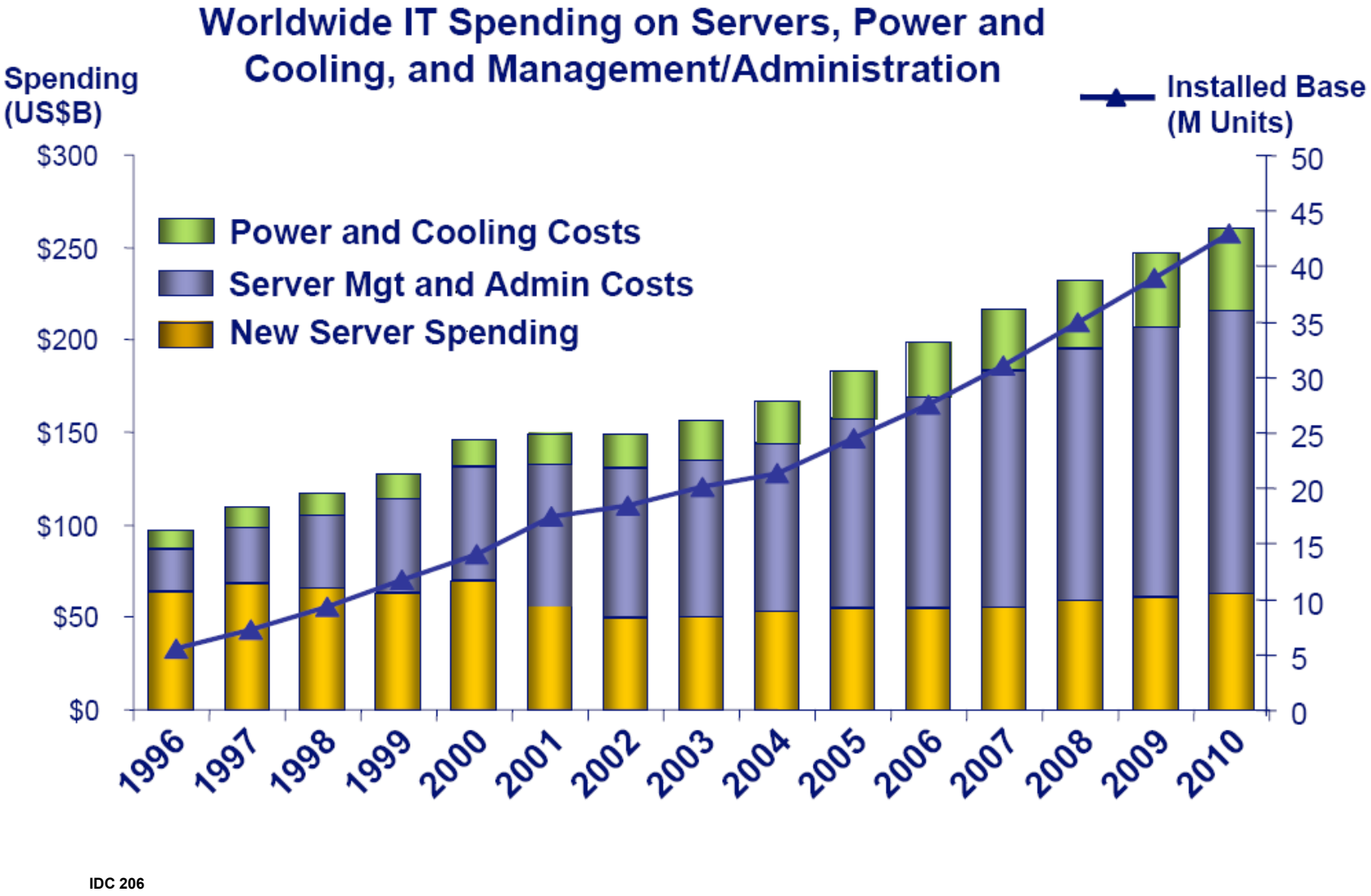



Legend
Java Program
J9LibOS and Support
9P Path
9P Component
Xen Dom0 Linux
Linux
GPFS Component
gpfs
Common Linux
Linux
Network Interfaces
9.2.70.236
DomU
DomU
DomU
Dom0
BE
BE
mgt
gpfs
mgt
gpfs
J9LibOS
J9LibOS
Linux
9p
Linux
SM
JS21 XenPPC
192.168.X.Y
DomU
DomU
DomU
Dom0
BE
BE
BE
mgt
gpfs
J9LibOS
J9LibOS
J9LibOS
9p
Linux
SM
JS21 XenPPC
DomU
DomU
DomU
Dom0
BE
BE
BE
mgt
gpfs
J9LibOS
J9LibOS
J9LibOS
9p
Linux
SM
JS21 XenPPC

Outline – Forces of Change

- **Virtualization**
- **Total Cost of Ownership (TCO)**
  - Scale-out / Cloud Computing / Web 2.0
  - Power budget
- **Hardware evolution**
  - Multicore
  - 3D integration/packaging
- **Bug Resistance**

Total Cost of Ownership (TCO)



IDC 206

Addressing TCO

- **Traditional datacenter workloads**
  - System mng, image mng, …
- **New workload trends**
  - Large clusters of commodity systems
    - Google, yahoo, ebay, amazon, MS, …
    - Failure happens and it will happen and be prepared!
  - Web 2.0 workloads:
    - One approach: optimize for simplified assumptions e.g. google:
      - GFS (SOSP'03)
      - MapReduce (OSDI'04) – Hadoop
      - BigTable (OSDI'06)
      - Chubby Lock Server (OSDI'06)

TCO-related (cont)

- **Amazon**
  - EC2 and S3
  - “shopping cart” functionality:
    - Traditional DB too heavy!
    - SOSP’07: distributed, peer-to-peer (p2p) approach
- **“Ressurgence” of Distributed Computing in Systems**
  - Byzantine algorithms
- **Our work on this:**
  - CSO: ICS’07, SNMP’07
  - A web-scale platform: ACM SIGOPS OSR Jan/08
- **Power management**

Multicore

- **Hardware evolution**
  - Weak cores, tons of them
- **Parallel programming to the masses!!**
  - It's so hard …
  - **view**.eecs.berkeley.edu/
- **DB has been successful in hiding concurrency complexities … why not us?**
- **Transaction Memory**
  - ASPLOS'07, ISCA
  - HotOS'07, SOSP'07
- **Our approach - accelerators**

Bugs

- **Static Analysis / Model checking**
  - Engler (SOSP/OSDI 2000-2006)
  - Coverity
- **YY Zhou at UUIC**
  - Using comments to pinpoint bugs (SOSP'07)
  - Bug allergies
  - Inferring multi-variable access (SOSP'07)
  - Comprehensive studies: disk (FAST'08), concurrency (ASPLOS'08)
- **Going beyond "core dumps" (SOSP'07, HotDep'07)**
- **Mining to identify deviation from "normal execution" (OSDI'06, SOSP'07)**

Conclusion

- **Other areas**
  - Browser as an OS
  - Will I have my digital pictures in 30 years?
  - File systems
- **Old/new problems … with exciting twist**
- **Best sources to follow what is going on:**
  - SOSP, OSDI, HotOS (every two years), Usenix
- **Surveys of industry work**
  - ACM SIGOPS special issues on:
    - Microsoft
    - IBM (January 2008)
    - HP