Dynamically Hacking the Kernel with Containers

ContainerCon Japan 2016
Tokyo

Quey-Liang Kao
National Tsing Hua University, Taiwan

About Myself

- Research topics
  - HPC (numerical), Heterogeneous computing
  - High-end hardware virtulization (InfiniBand, GPGPU)
  - Container technology
- Contributions
  - AUR packager and maintainer
    - runc-git, openscap
    - kpatch

Outline

- Motivation
- Background
- Demo: kernel detouring
  - FreeBSD on Linux
- Other approaches
- Conclusion

Motivation

http://typemoon.wikia.com/wiki/Holy_Grail

Higher
Performance

VM

Container

Lower
Isolation

Quick Survey

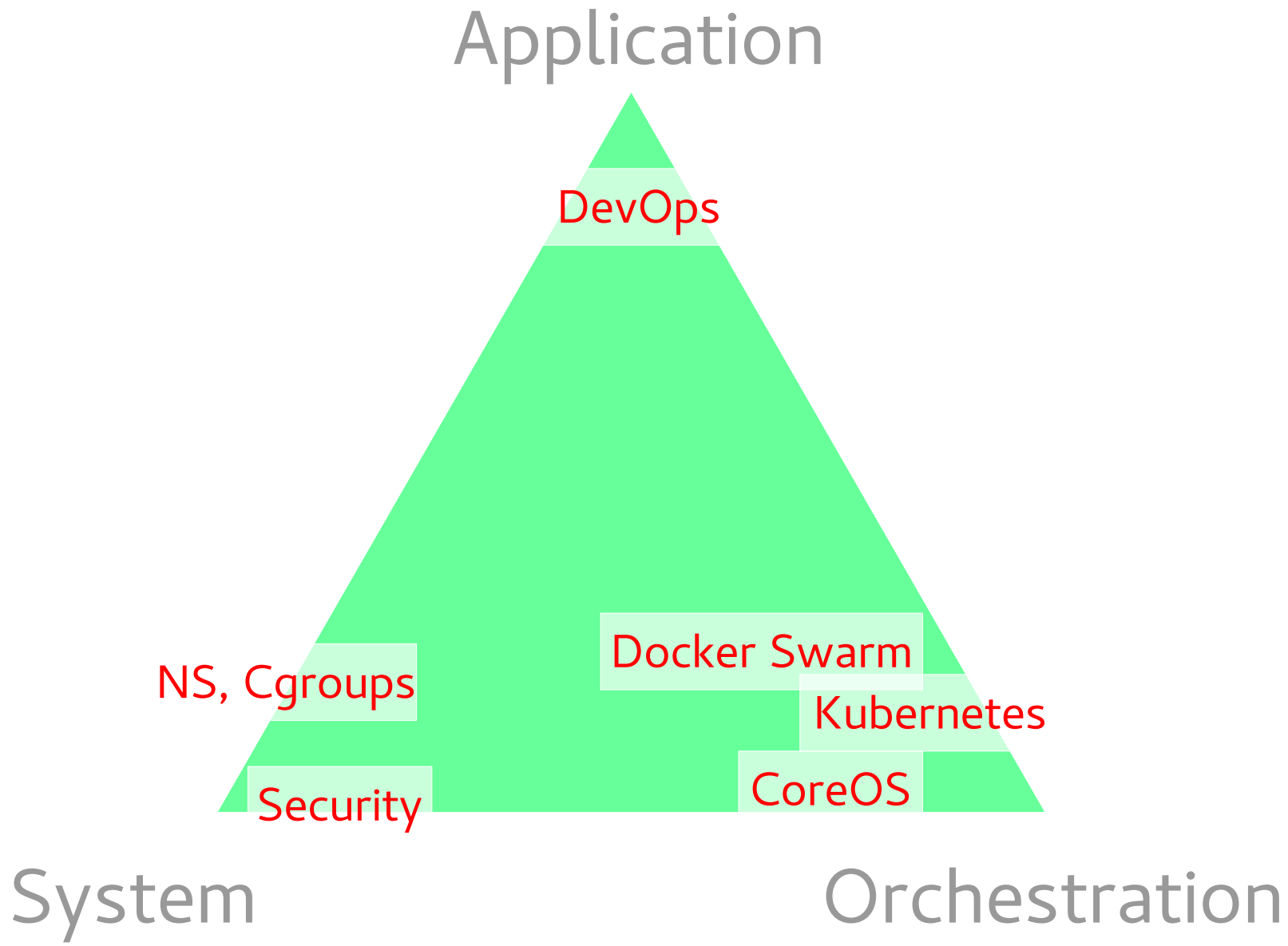
Terms

- OS container
  - Like a VM
  - No layered FS
  - LXC, OpenVZ, BSD Jails, Solaris zones
- App container
  - For single service
  - layered FS
  - Docker, Rocket

https://blog.risingstack.com/operating-system-containers-vs-application-containers/

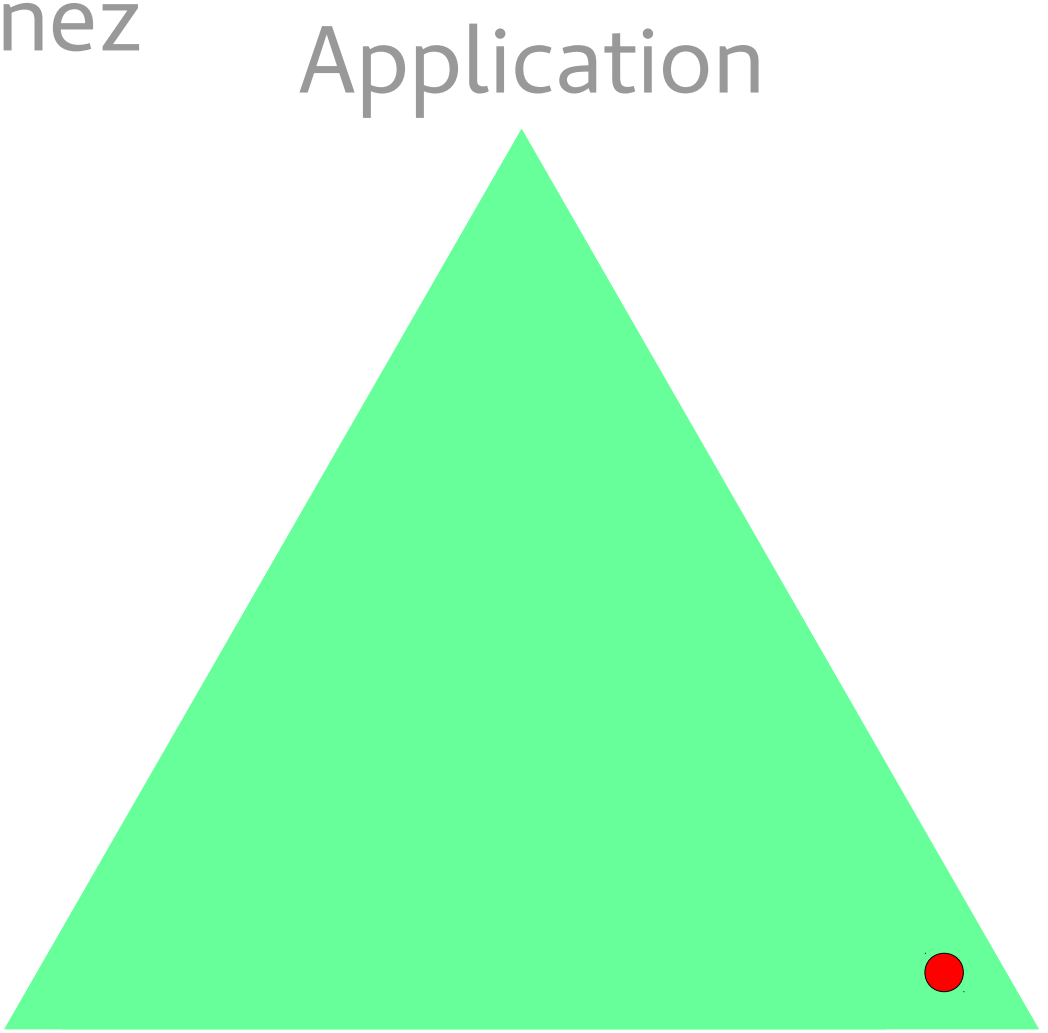
There Dimensions

- **System software**
  - OS container's perspective
- **Orchestration**
  - App container's perspective
- **Applications**


Application
DevOps
Docker Swarm
NS, Cgroups
Kubernetes
CoreOS
Security
System
Orchestration

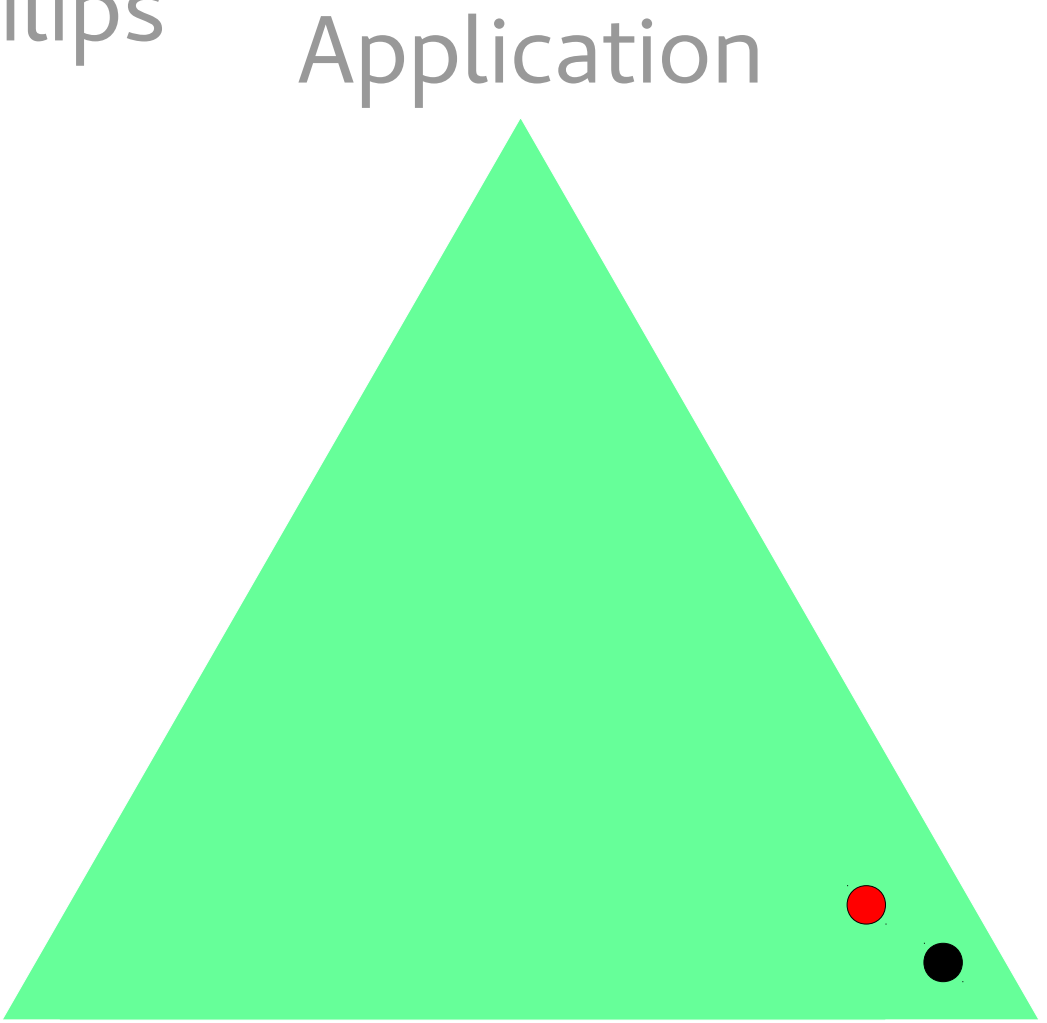
How do I Orchestrate My Container?

By Isabel Jimenez



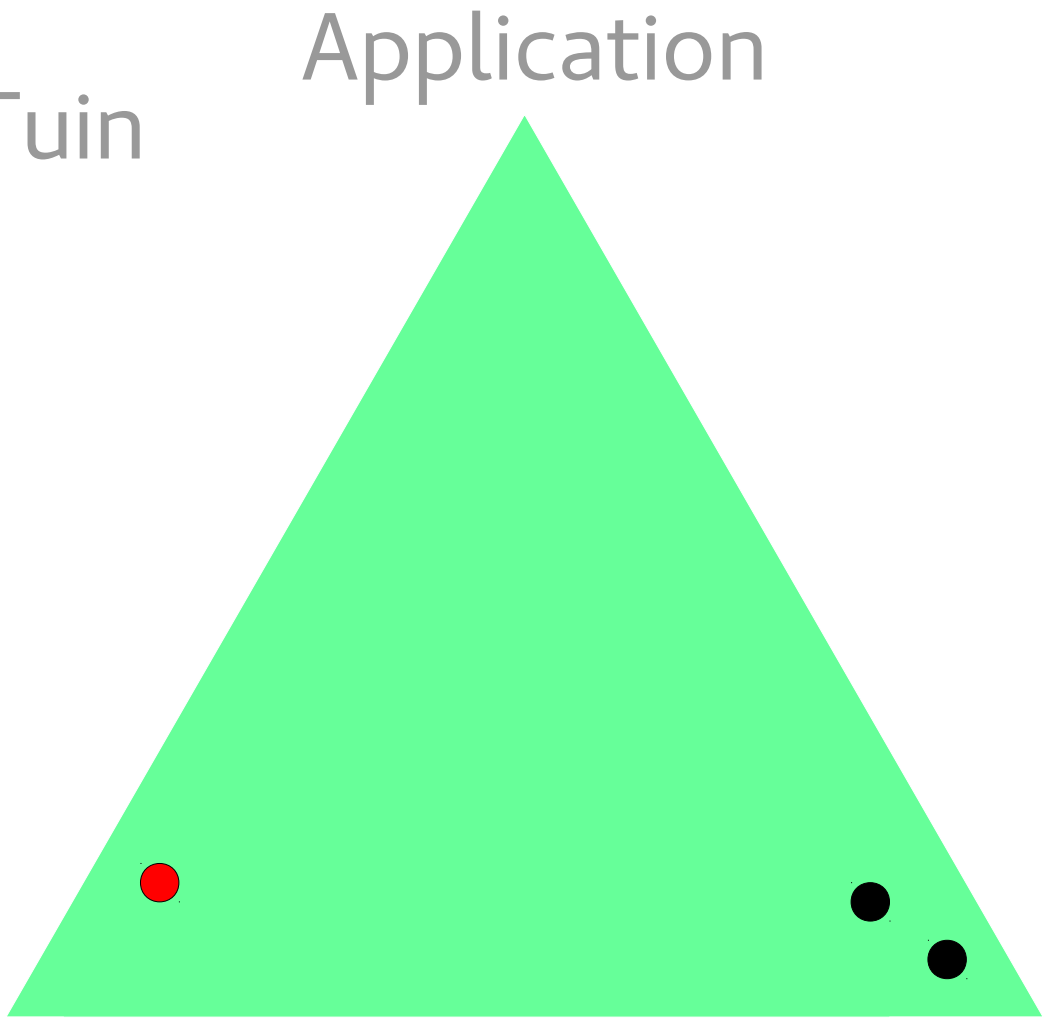
Containers at scale thanks to Kubernetes

By Brandon Philips



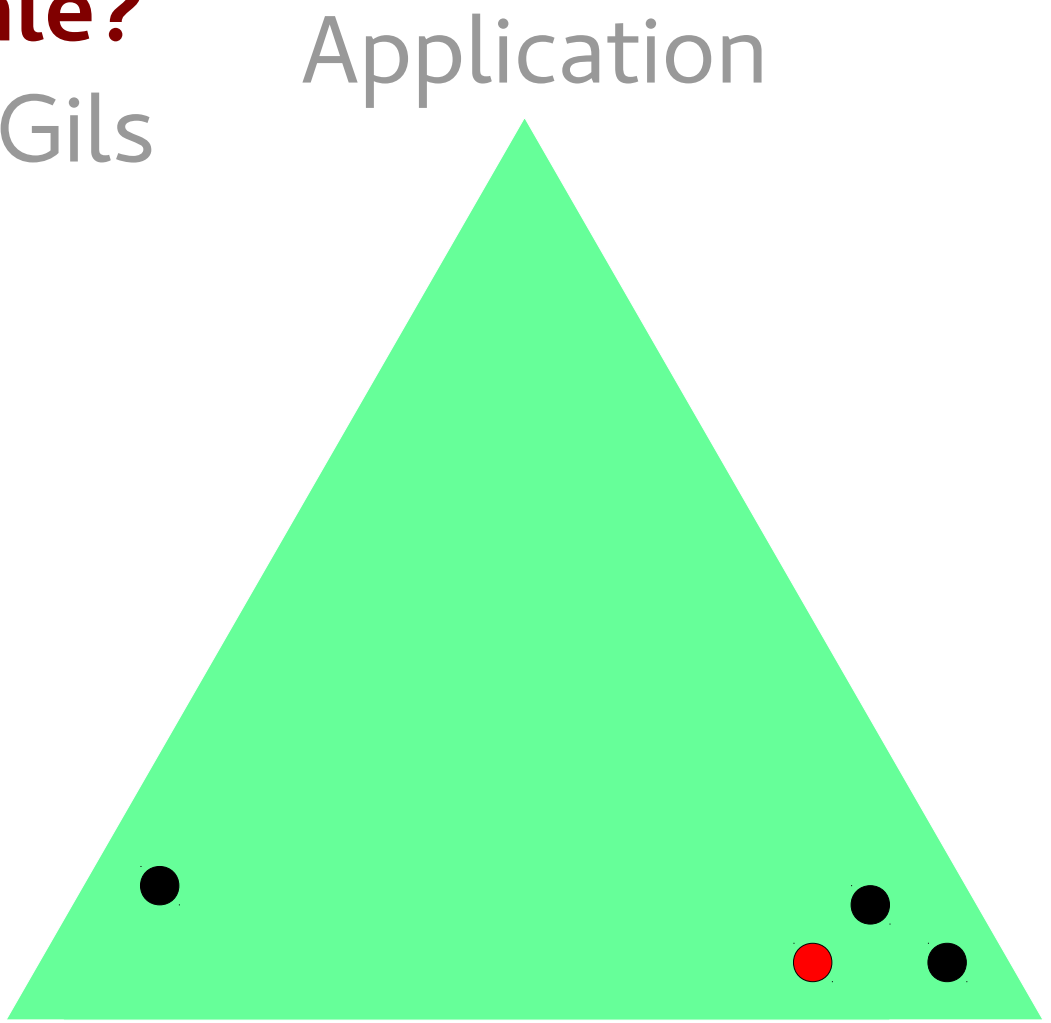
A Security State of Mind: Container Security in the Enterprise

By Chris Van Tuin



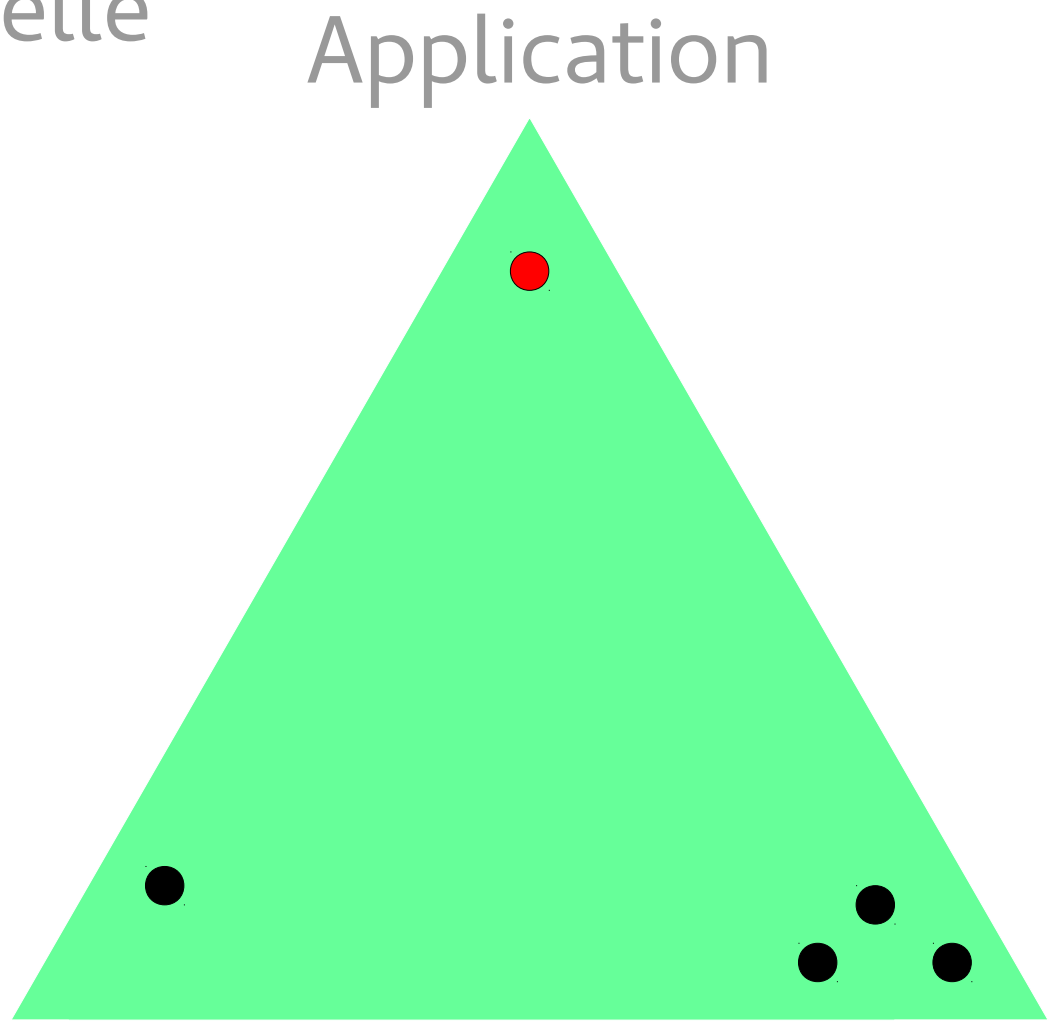
How the hell do I run my containers in production, and will it scale?

By Daniël van Gils



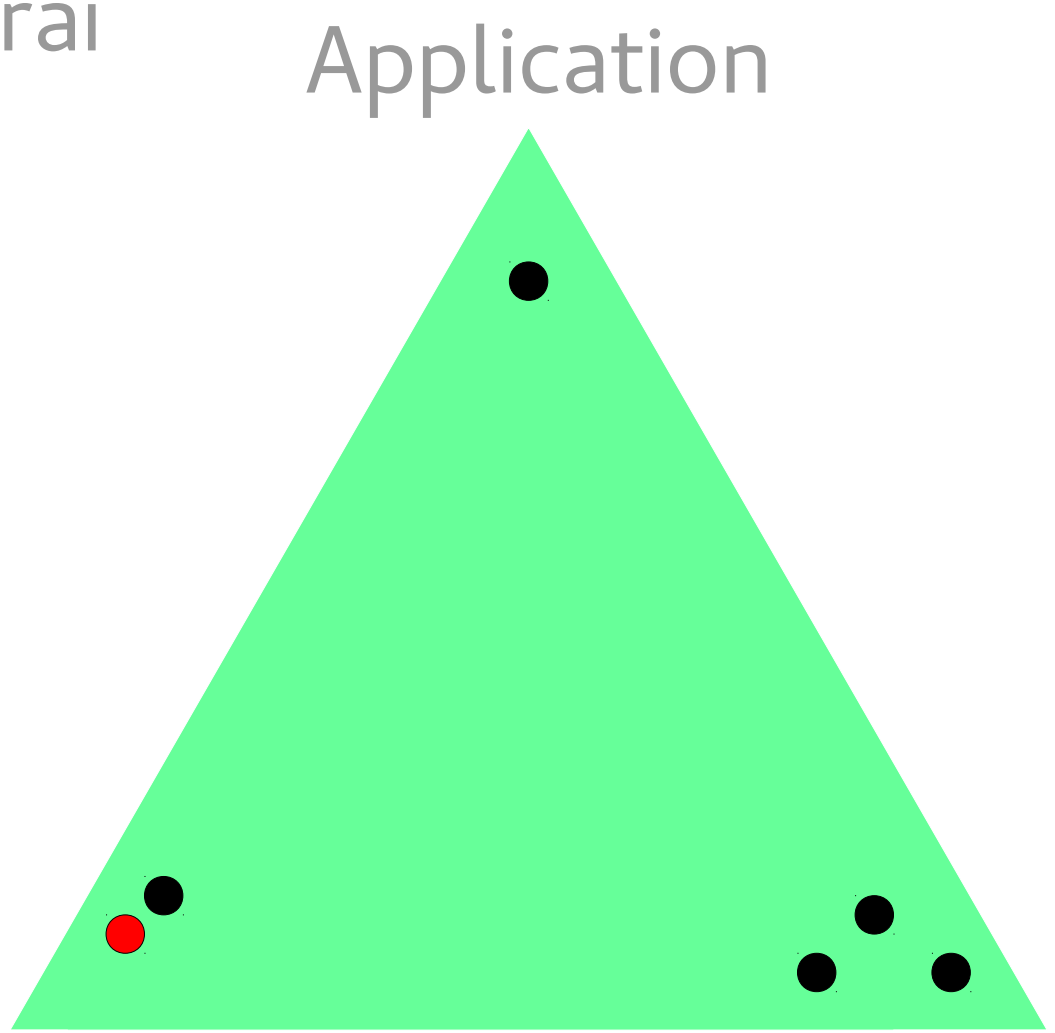
Container Hacks & Fun Images

By Jessie Frazelle



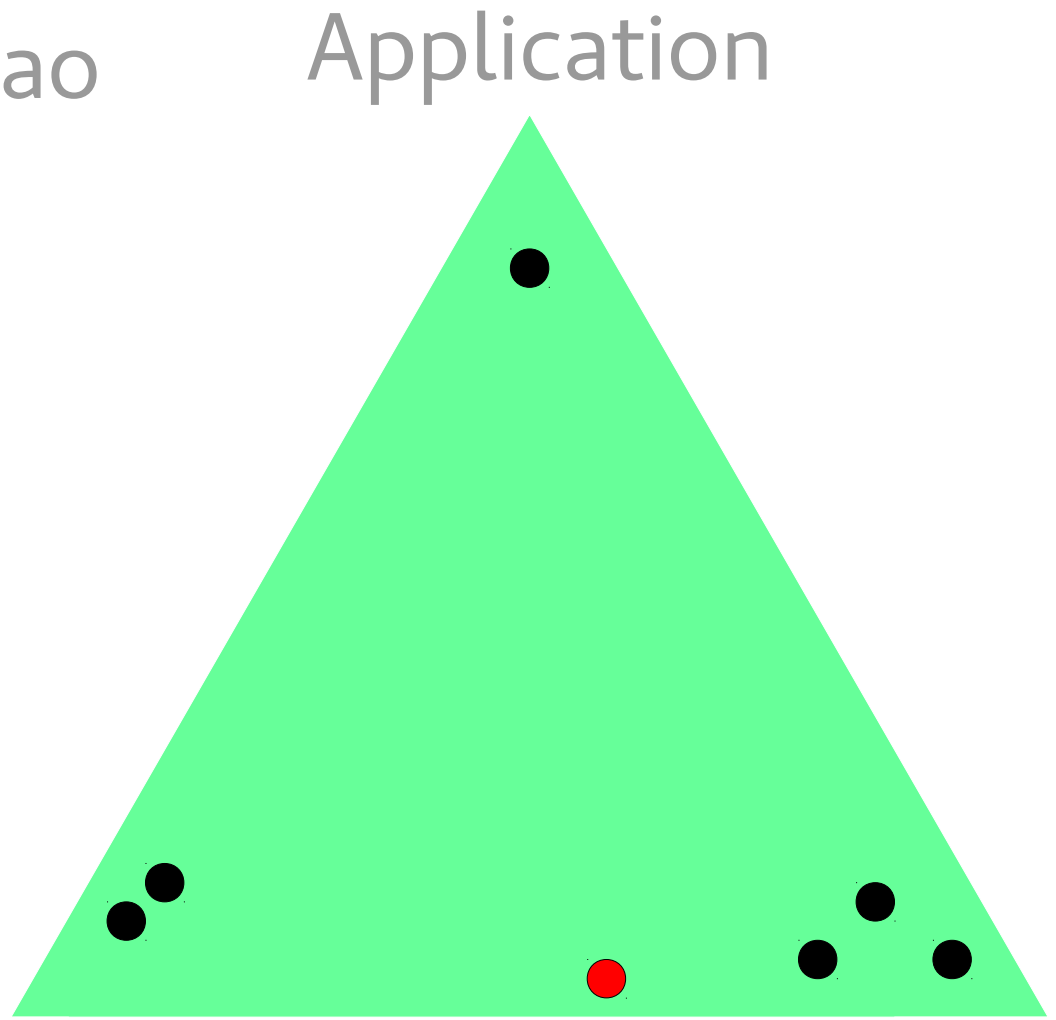
Rootless Container with Runc

By Aleksa Sarai



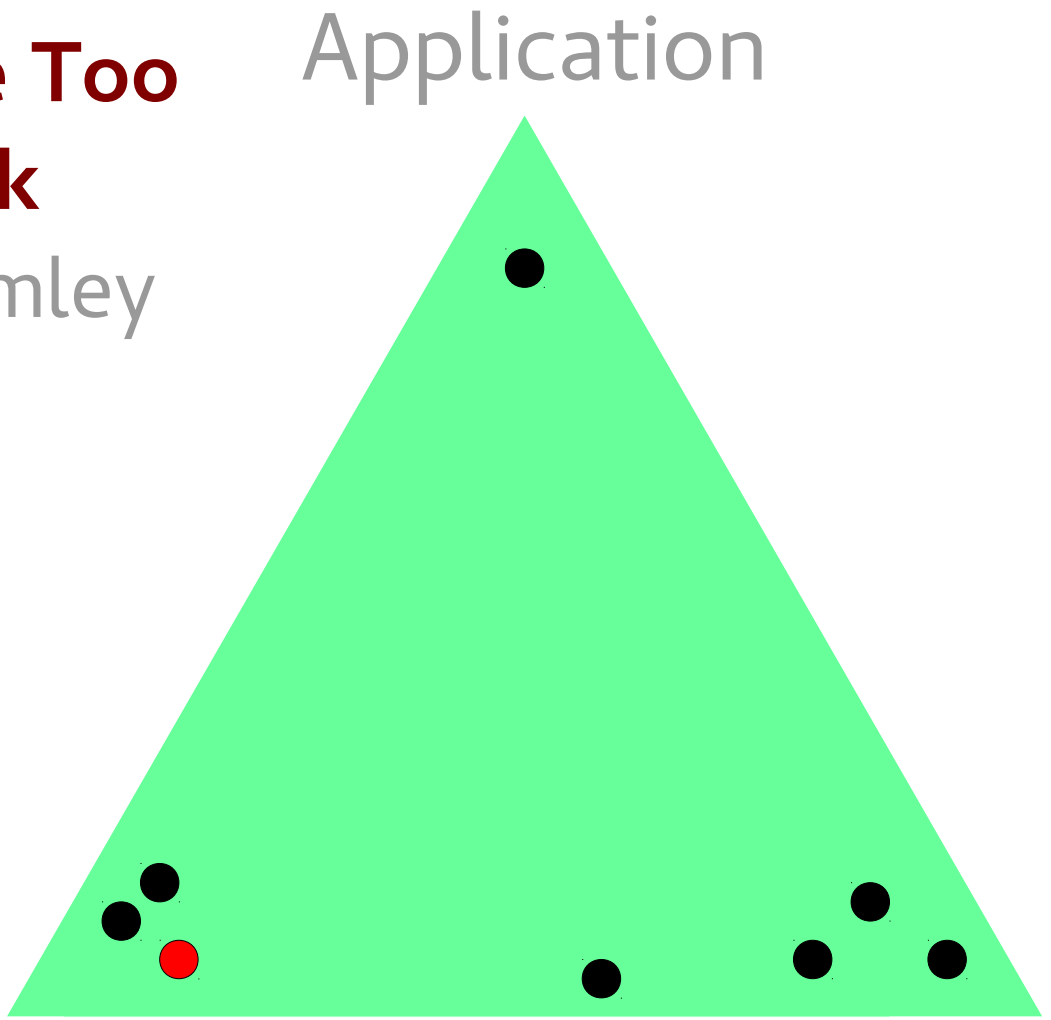
Soft Container Towards 100% Resource Utilization

By Accela Zhao

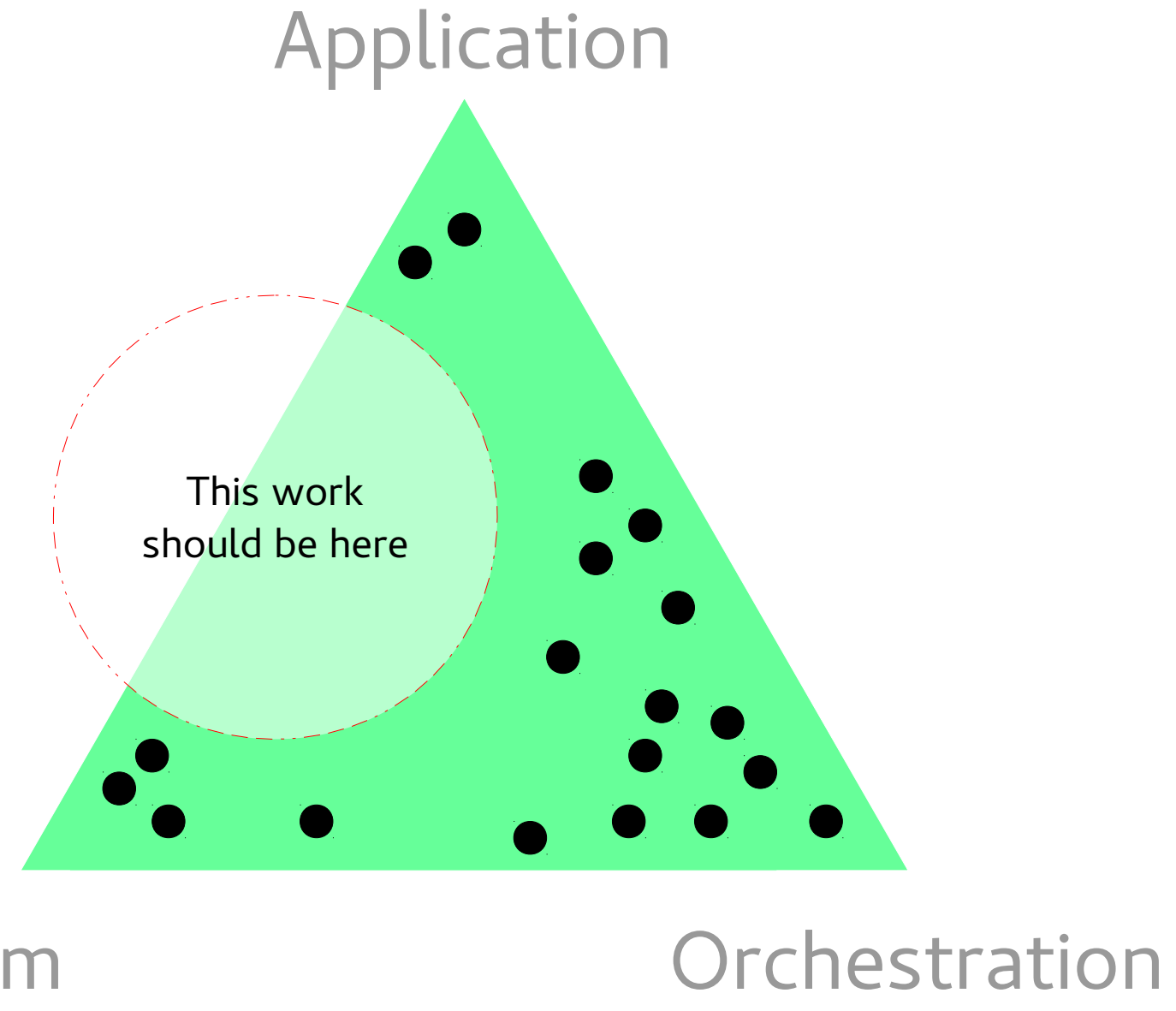


Unprivileged Containers: What you Always Wanted to Know About Name-spaces But Were Too Afraid To Ask

By James Bottomley



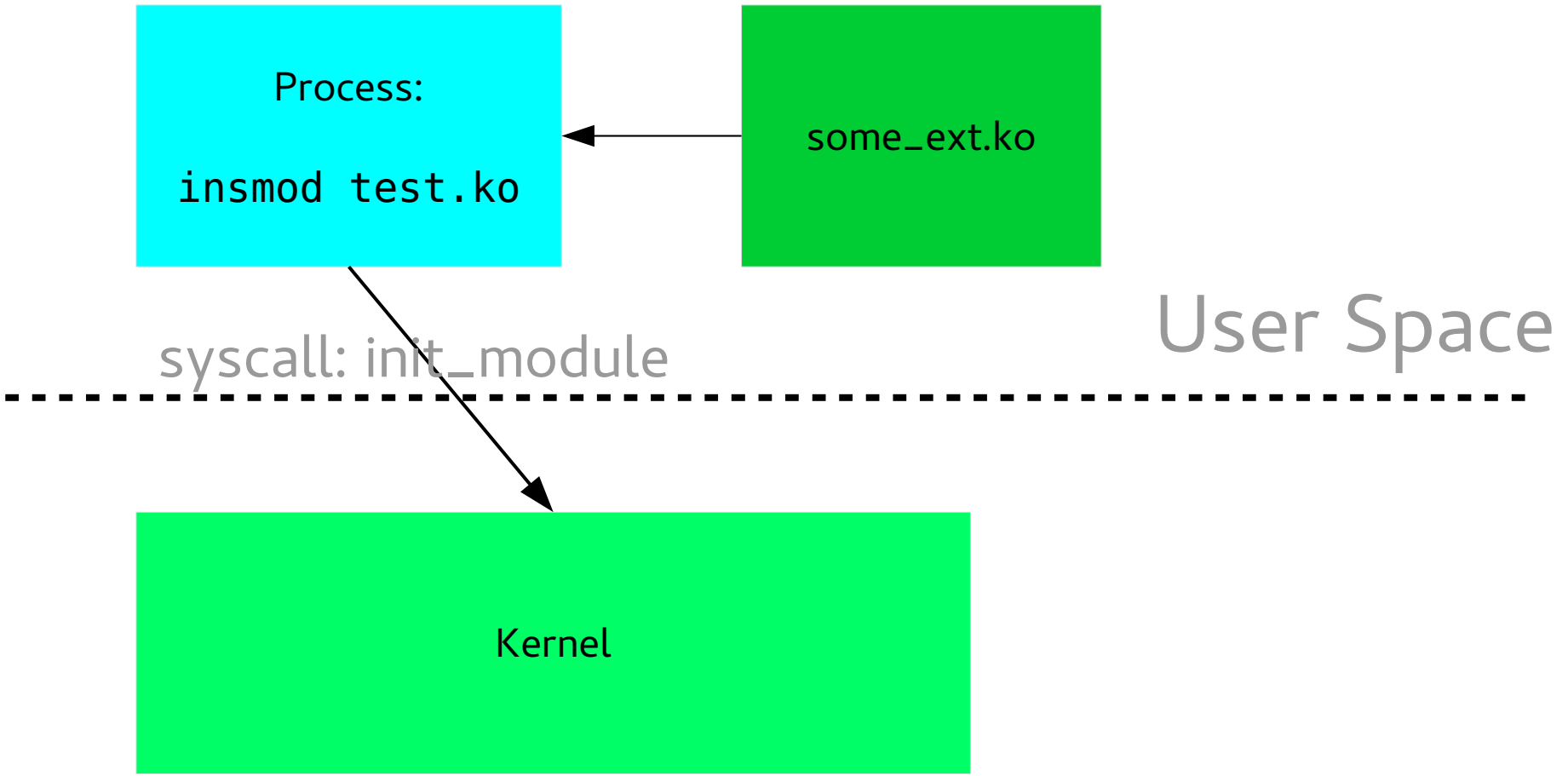
etc...


Application
This work
should be here
System
Orchestration

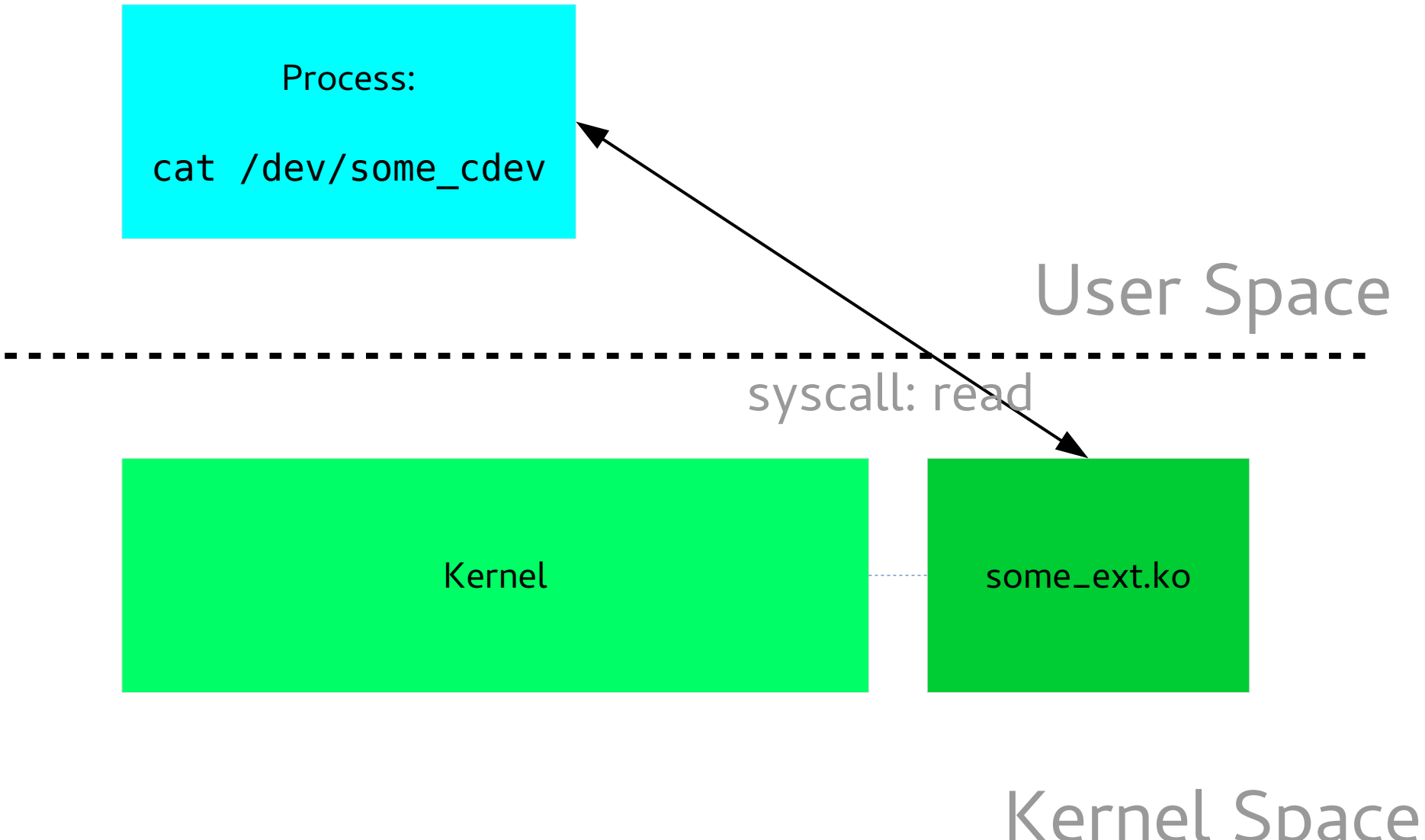
Background

Modules,
Live patches,
and **Kerenl detouring**

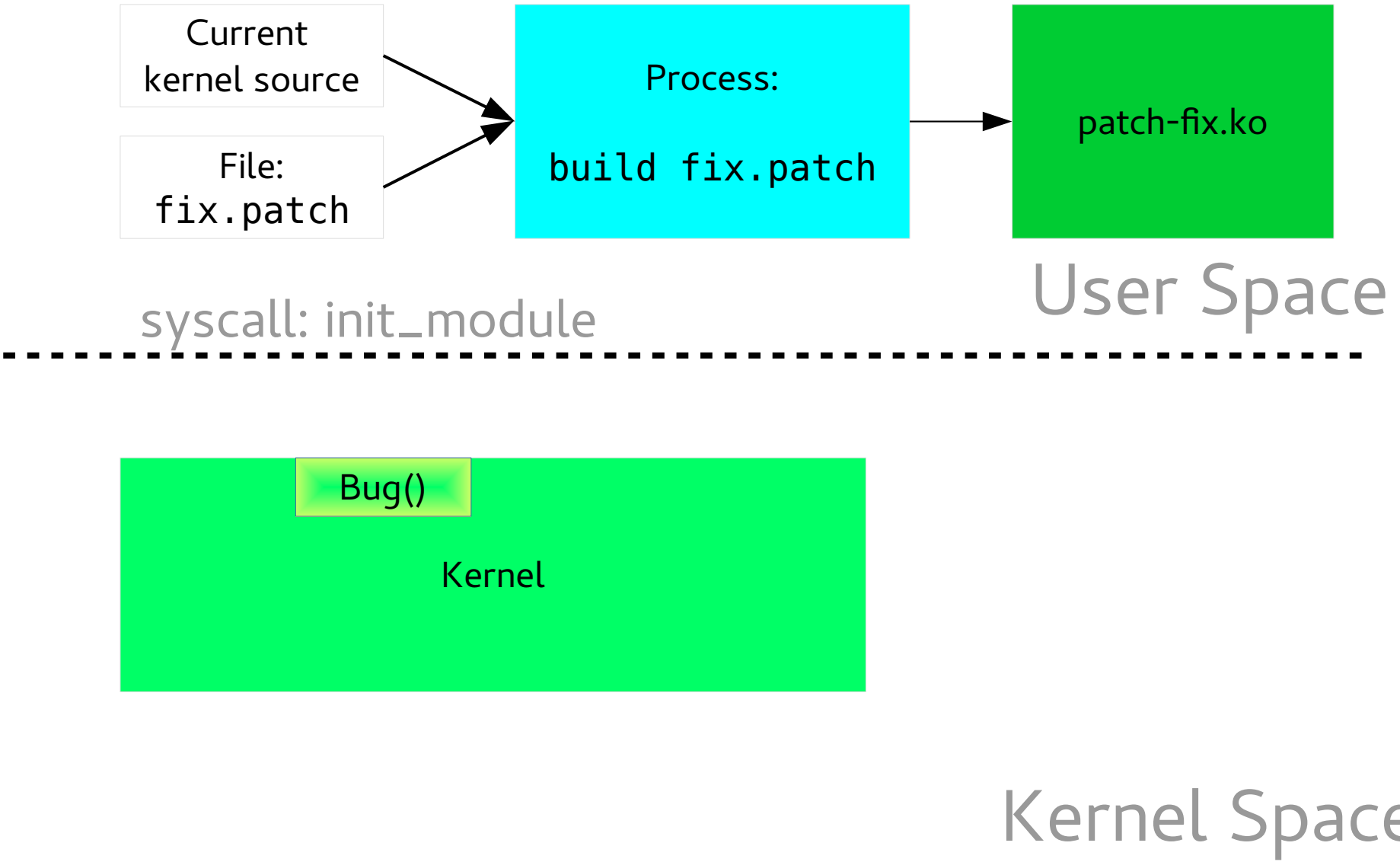
Kernel Module: Loading



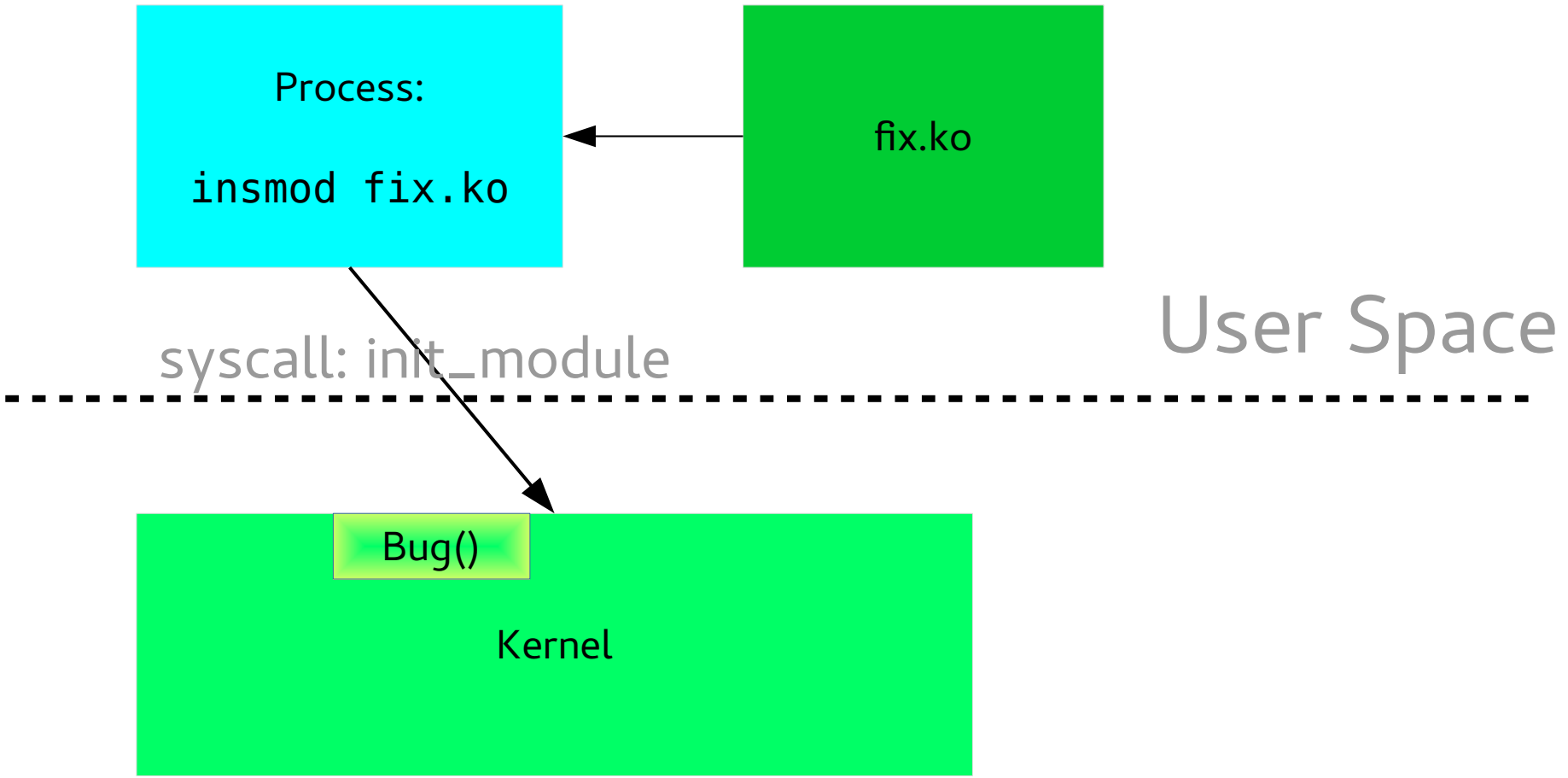
Kernel Module: Using



Live Patching: Building

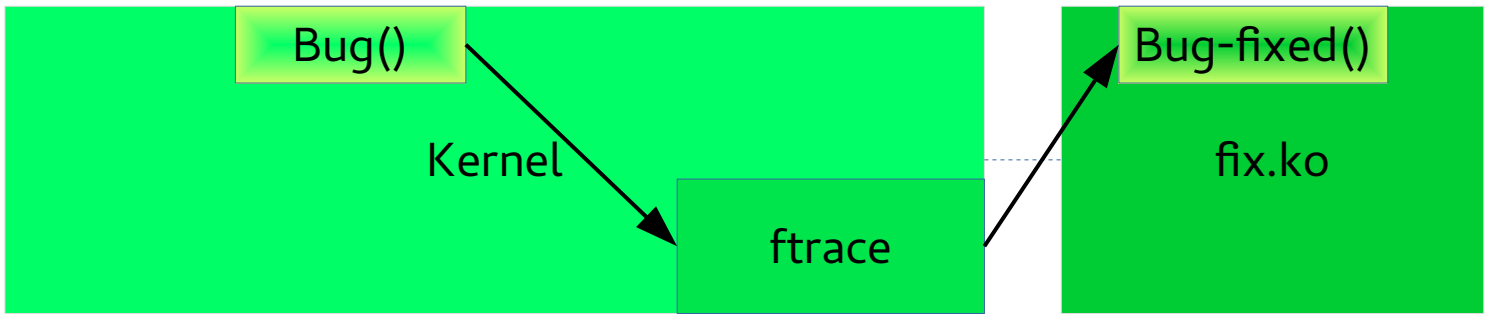


Live Patching: Applying



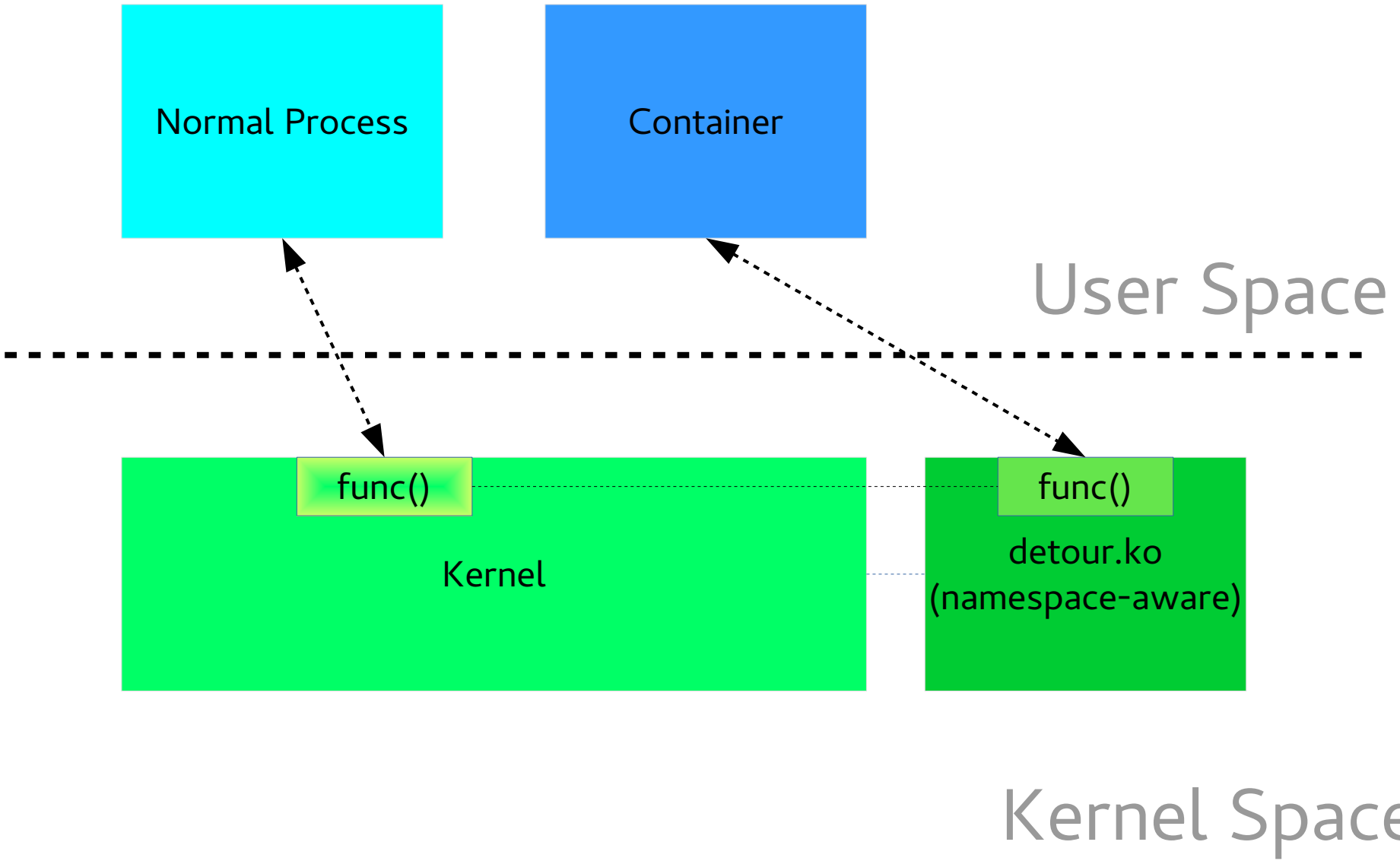
Live Patching: Applying

User Space



Kernel Space

Kernel Detouring

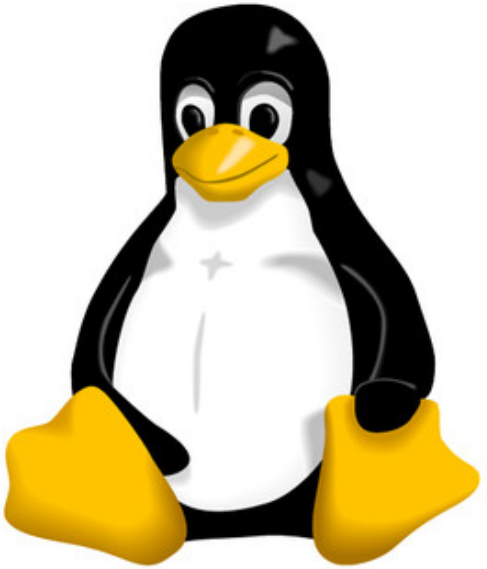


Demo: Kernel Detouring

http://kirokueiga.seesaa.net/archives/201208-1.html

FreeBSD binary on Linux

```
01000110011100100110010101100101010000100101001101000100
```



Specific Challenges (FreeBSD)

- Corresponding system calls
  - Flag numbers are not portable
  - different calling/exiting conventions
- Unique system calls
  - Re-implementation

General Challenges

- Insufficient isolation
- Limitation of development
  - live patching should only be a temp. solution

Other Binary Compatibility Work

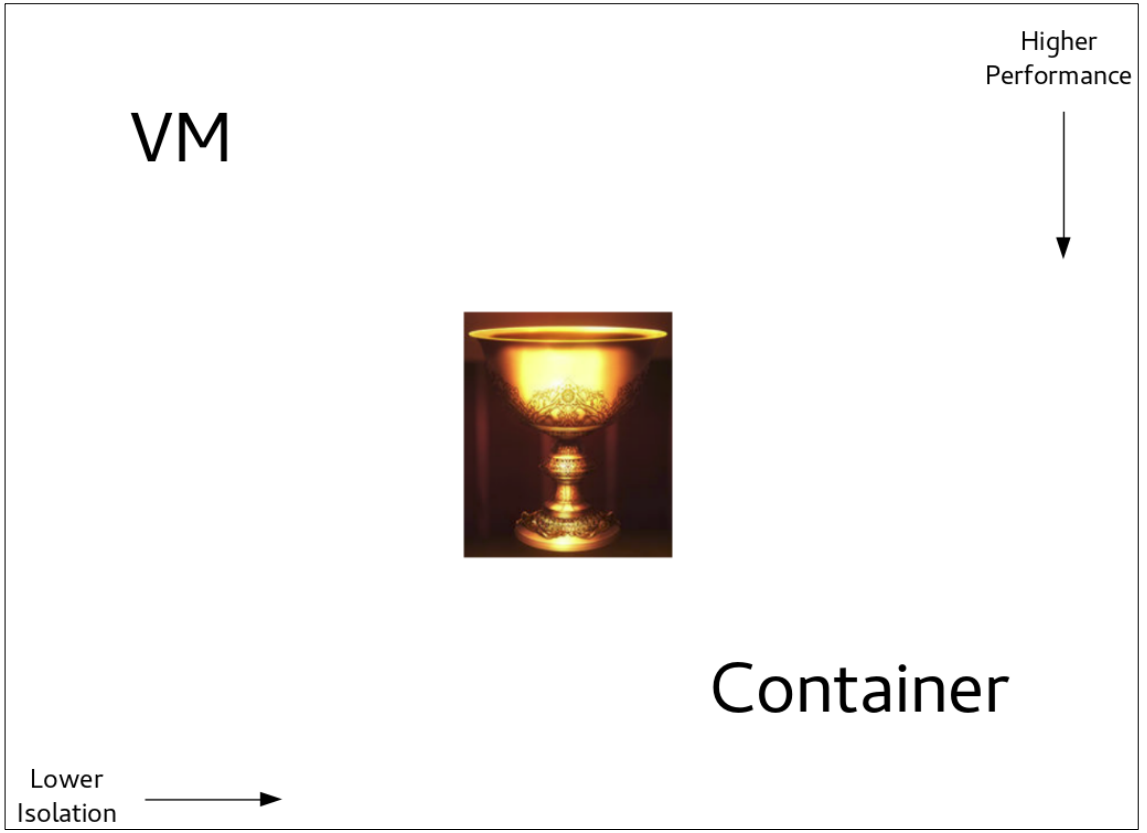
- Wine
  - Special loader for PEs/DLLs
- FreeBSD, Windows 10
  - Kernel built-in compatibility layer for Linux binary
  - System call remapping/re-implementation

Possible Applications

- Experimental module/patch test bed
- Images for other OSes
- Educational purpose
  - why not?

Other approaches

- Hyper-V
- Multi-Kernel
- Unikernel

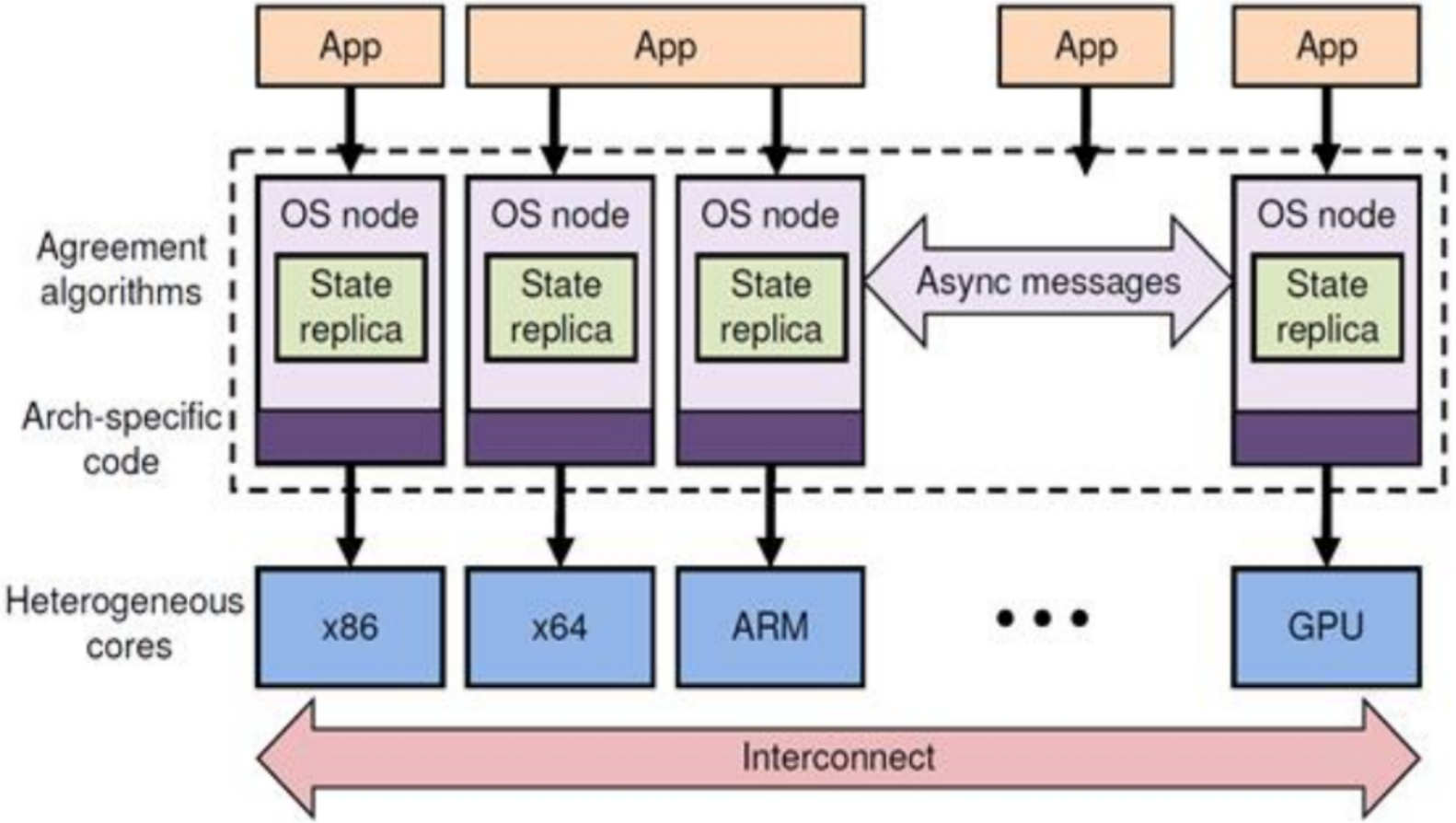


Microsoft Hyper-V

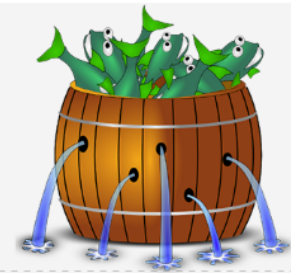
- A private kernel per container
  - stripped kernel reduced from Windows server
- Unix-likes support
  - as VMs (in a VM-like container) ~~(container-like VM)~~

Multi-Kernel

- Barrelfish
  - Philosophy: **separation** and **duplication** rather than keep syncing
  - One kernel per core
  - Scalability and heterogeneity
- VirtuOS, Arrakis, Quest-V, etc.
  - Performance improvement

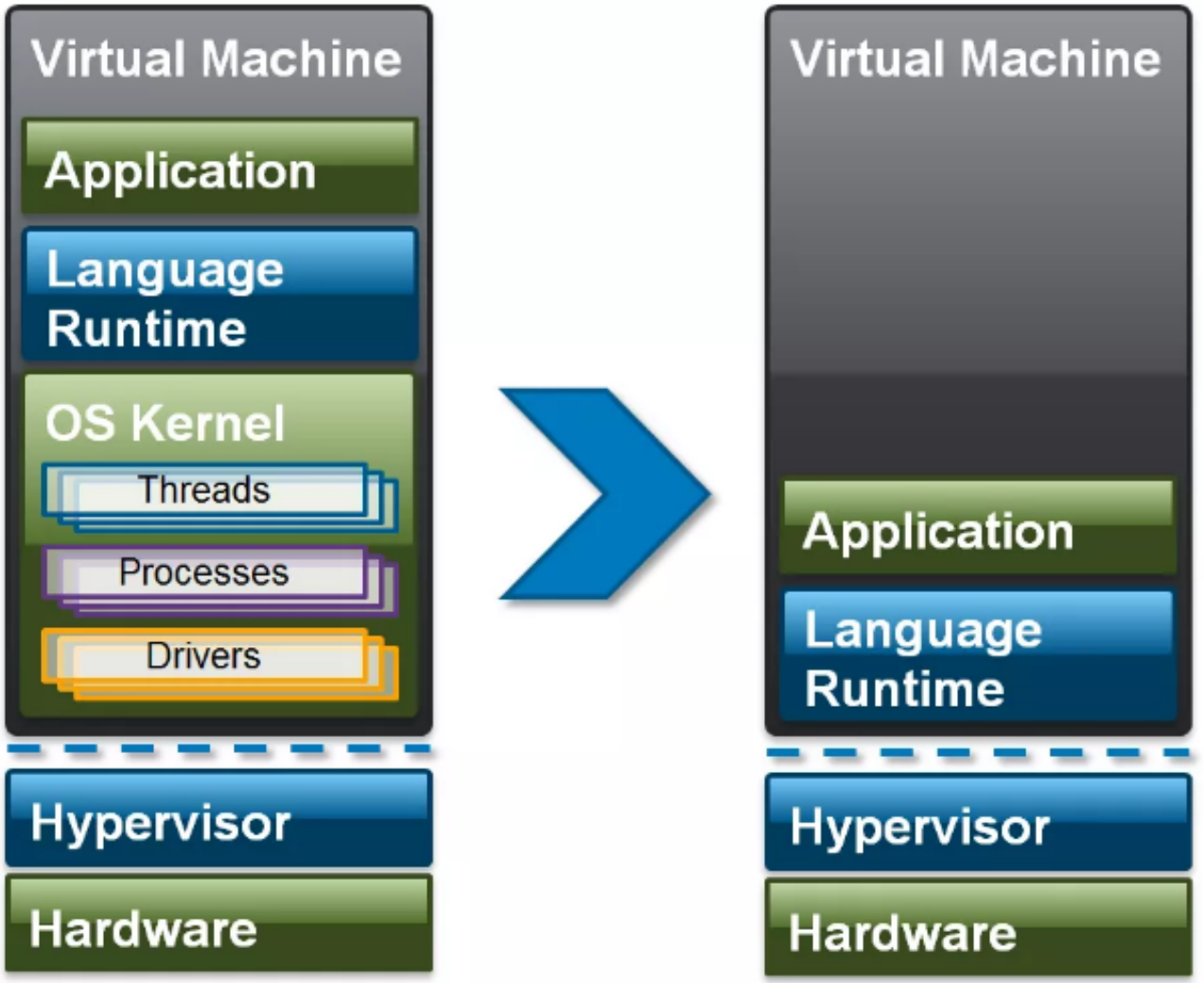


http://courses.cs.washington.edu/courses/csep551/14au/video/archive/html5/video.html?id=csep551_14au_6



UniKernel

- Rump Kernel, MirageOS, OSv, etc.
  - Application oriented
    - no more general-purpose
  - “Compiler” approach



http://www.penninkhof.com/2015/05/minimalist-cassandra-vm-using-osv/

http://typemoon.wikia.com/wiki/Holy_Grail

Higher
Performance

VM

Hyper-V

New paradigms
(Multi-kernel, Unikernels)

Kernel detouring
(or some kernel-space ext.)

Lower
Isolation

Container

Conclusion

- The **kernel detouring** demo attempts to indicate a possible movement of the development of OS containers
  - as a proof-of-concept
- Future direction
  - Make more fun
  - Make it more complete

Q & A