KubeCon

CloudNativeCon

Europe 2018

Low-overhead tracing using eBPF

Gaurav Gupta, SAP Labs

Gaurav Gupta

Developer @ 

gaurav.gupta07@sap.com

Agenda

KubeCon | CloudNativeCon Europe 2018

Mission:

- Apply modern, low-overhead, production-ready BPF-based tools for performance investigations and monitoring in Linux systems

Objectives:

- Understand the advantages of BPF-based tracing tools
- Trace system events in real-time with BCC
- Generate stack traces for system events
- Develop on-demand performance tools with BCC

eBPF

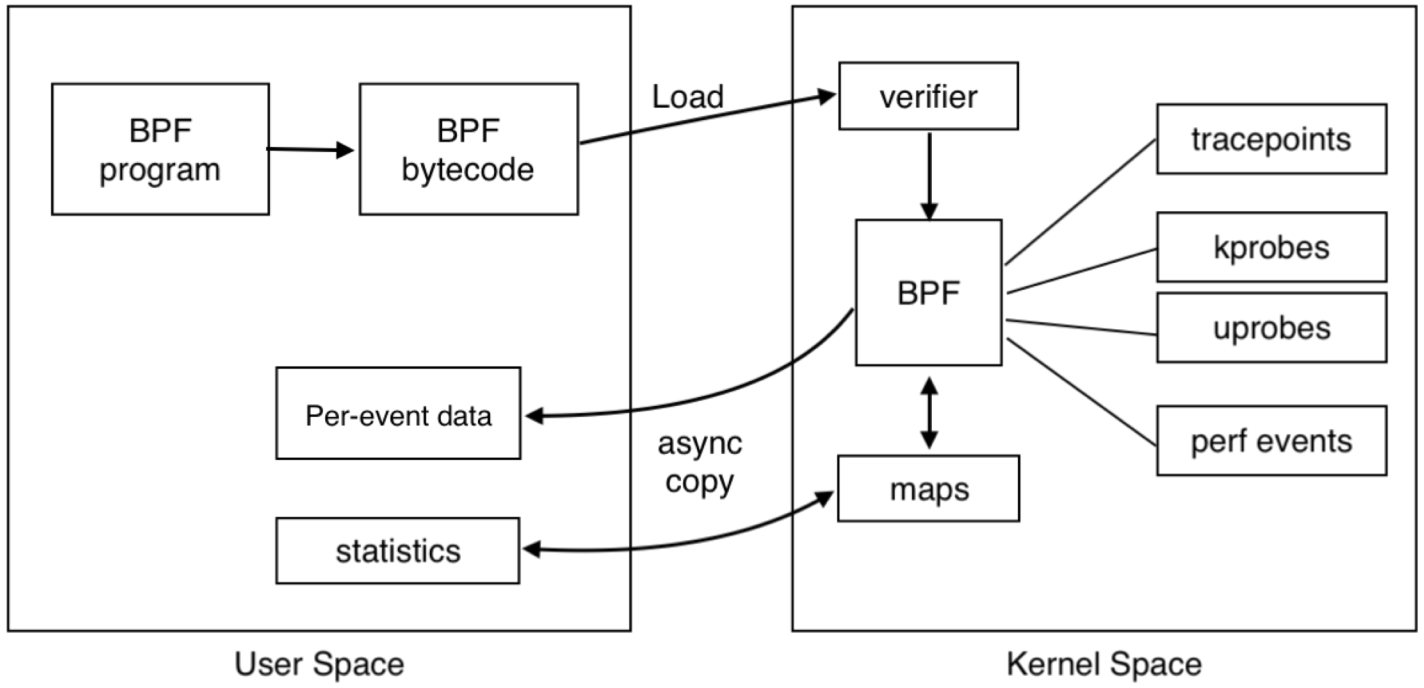
"The next Linux superpower"
- Brendan Gregg

eBPF (extended Berkeley Packet Filter)

- Functional in Linux 3.19
- Enhanced in 4.x kernels

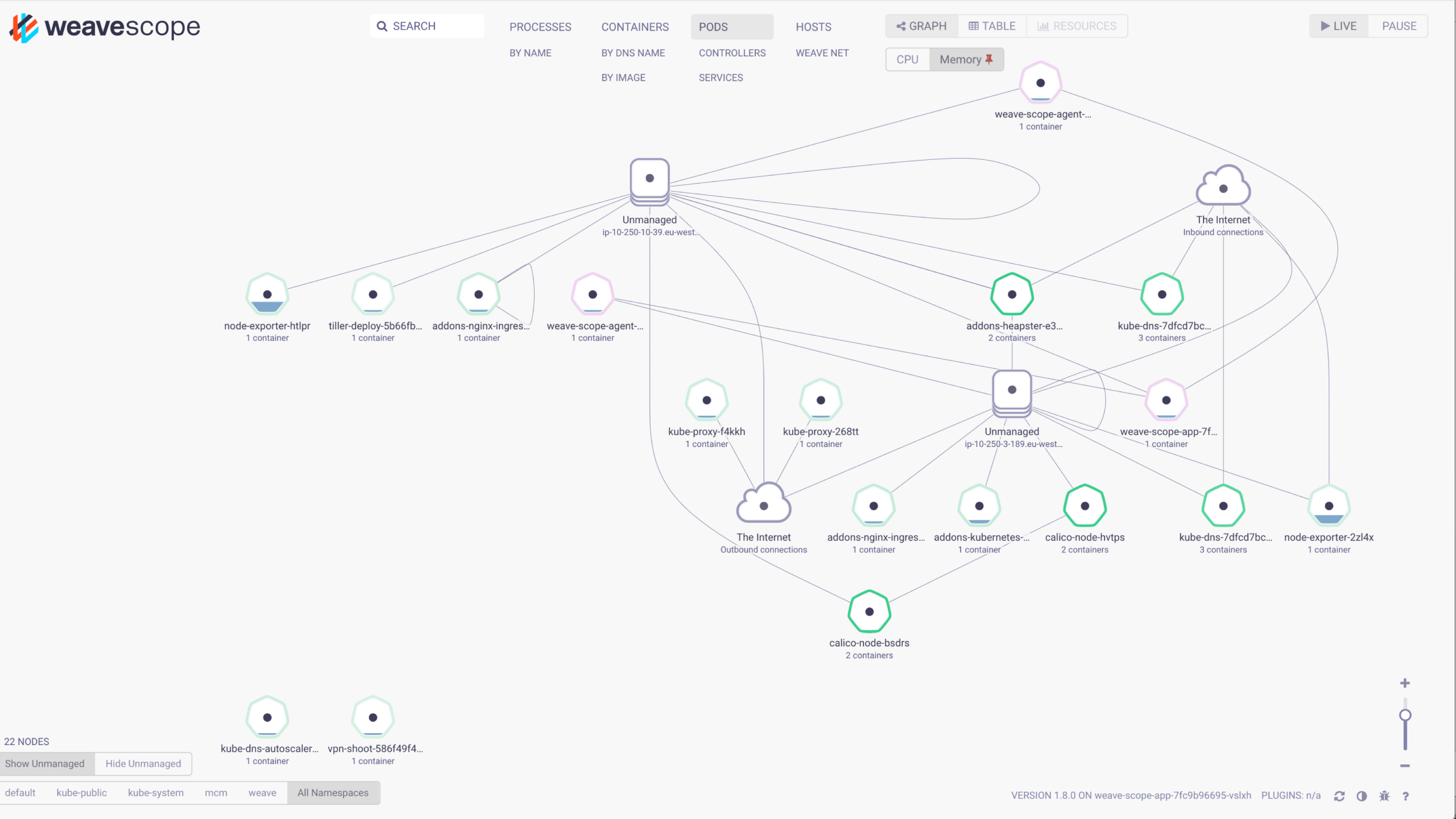
Workflow

- eBPF Program → Bytecode
- Verifier: checks validity
- Load into BPF machine
- Trace user/kernel code
- eBPF maps



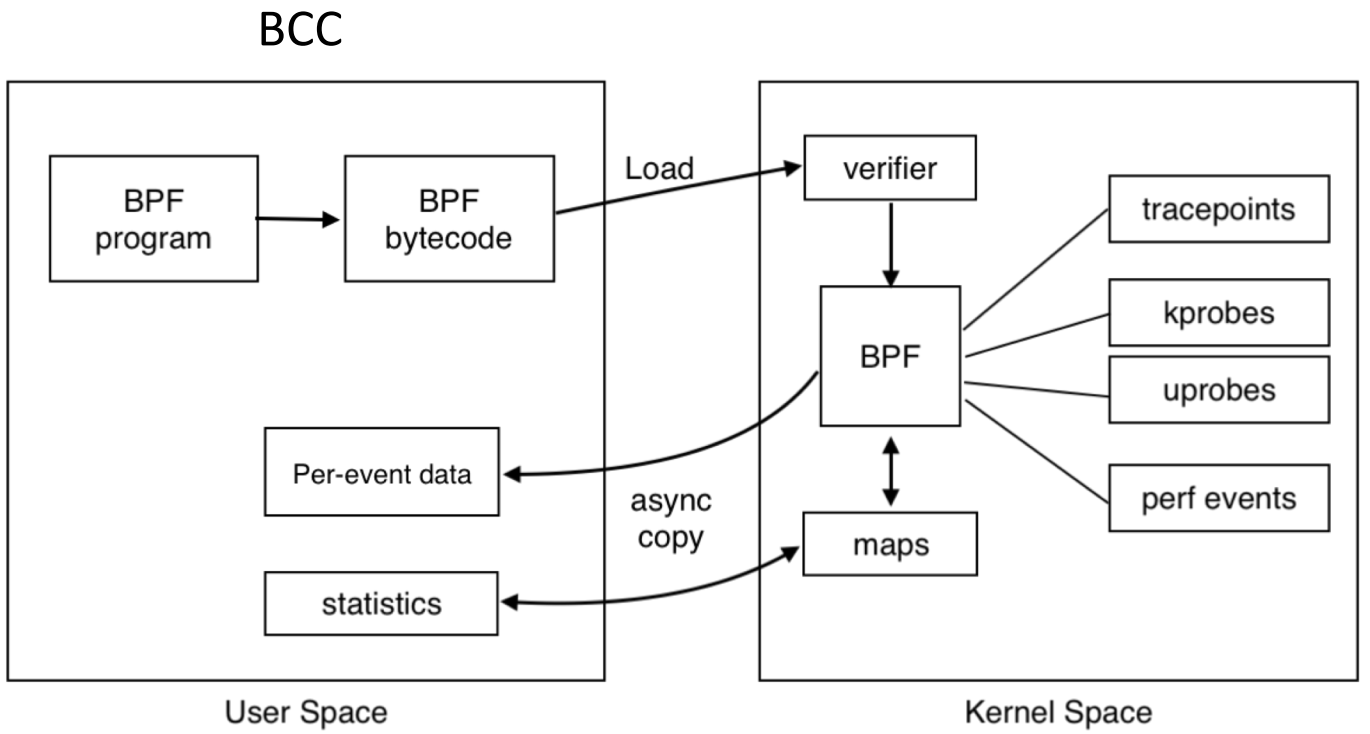
Tools based on eBPF

- Seccomp
  - Allow 3rd party code in a safer manner
- Cilium
  - Provides networking, security, load-balancing and visibility for application containers
- eBPF in Kubernetets
  - https://kubernetes.io/blog/2017/12/using-ebpf-in-kubernetes


weavescope
SEARCH
PROCESSES
BY NAME
CONTAINERS
BY DNS NAME
BY IMAGE
PODS
CONTROLLERS
SERVICES
HOSTS
WEAVE NET
GRAPH
TABLE
RESOURCES
LIVE
PAUSE
CPU
Memory
weave-scope-agent-...
1 container
Unmanaged
ip-10-250-10-39.eu-west...
The Internet
Inbound connections
node-exporter-htlpr
1 container
tiller-deploy-5b66fb...
1 container
addons-nginx-ingres...
1 container
weave-scope-agent-...
1 container
addons-heapster-e3...
2 containers
kube-dns-7dfcd7bc...
3 containers
kube-proxy-f4kkh
1 container
kube-proxy-268tt
1 container
Unmanaged
ip-10-250-3-189.eu-west...
weave-scope-app-7f...
1 container
The Internet
Outbound connections
addons-nginx-ingres...
1 container
addons-kubernetes-...
1 container
calico-node-hvtps
2 containers
kube-dns-7dfcd7bc...
3 containers
node-exporter-2zl4x
1 container
calico-node-bsdrs
2 containers
kube-dns-autoscaler...
1 container
vpn-shoot-586f49f4...
1 container
22 NODES
Show Unmanaged
Hide Unmanaged
default
kube-public
kube-system
mcm
weave
All Namespaces
VERSION 1.8.0 ON weave-scope-app-7fc9b96695-vslxh
PLUGINS: n/a

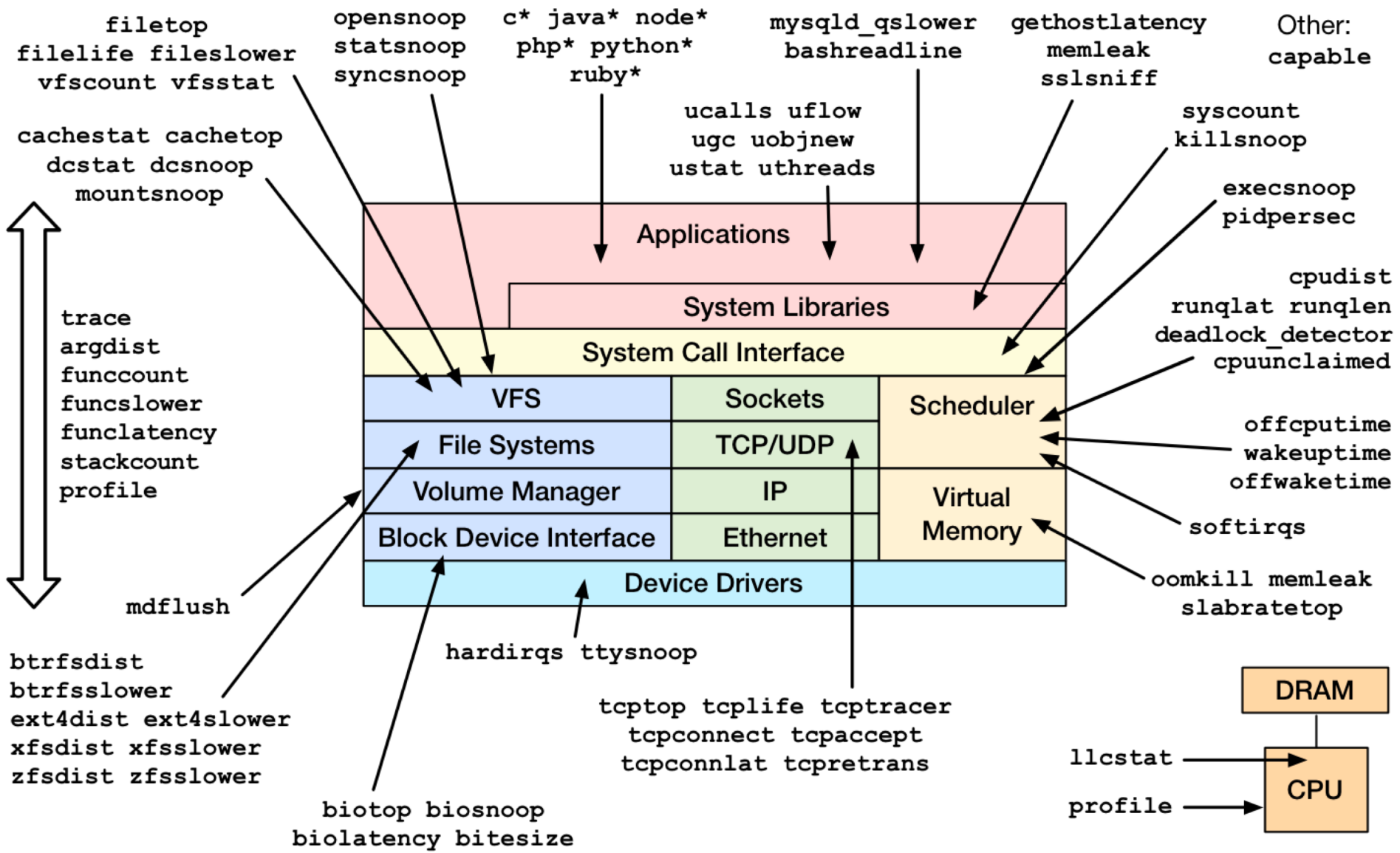
BCC

- BPF Compiler Collection (BCC) is a BPF frontend library
- Helps build BPF-based tools in high-level languages
  - Python, Lua, C++



https://github.com/iovisor/bcc

Linux bcc/BPF Tracing Tools



https://github.com/iovisor/bcc#tools 2017

Use cases

- Trace new processes
- Monitor tty or pts devices

- Observability
  - Function latencies
  - Monitor Hardware and software events
  - On-CPU and off-CPU profiles; low overhead

BCC Code snippets

Trace Kernel's `vfs_read` Latency

```
# Attach probes
b.attach_kprobe(event=vfs_read, fn_name="trace_func_entry")
b.attach_kretprobe(event=vfs_read, fn_name="trace_func_return")
```

```
int trace_func_entry(struct pt_regs *ctx)
{
    u32 pid = bpf_get_current_pid_tgid();
    u64 ts = bpf_ktime_get_ns();

    FILTER
    ENTRYSTORE
    start.update(&pid, &ts);

    return 0;
}
```

```
int trace_func_return(struct pt_regs *ctx)
{
    u64 *tsp, delta;
    u32 pid = bpf_get_current_pid_tgid();

    // calculate delta time
    tsp = start.lookup(&pid);
    if (tsp == 0) {
        return 0;   // missed start
    }
    delta = bpf_ktime_get_ns() - *tsp;
    start.delete(&pid);
    FACTOR

    // store as histogram
    STORE

    return 0;
}
```

bcc tools/funclatency.py

Example (1/3)
vfs_read latency histogram


```
Tracing 1 functions for "vfs_read"... Hit Ctrl-C to end.
^C
     nsecs               : count     distribution
         0 -> 1          : 0        |                                        |
         2 -> 3          : 0        |                                        |
         4 -> 7          : 0        |                                        |
         8 -> 15         : 0        |                                        |
        16 -> 31         : 0        |                                        |
        32 -> 63         : 0        |                                        |
        64 -> 127        : 0        |                                        |
       128 -> 255        : 0        |                                        |
       256 -> 511        : 84       |***************                         |
       512 -> 1023       : 229      |****************************************|
      1024 -> 2047       : 44       |********                                |
      2048 -> 4095       : 176      |******************************          |
      4096 -> 8191       : 43       |*******                                 |
      8192 -> 16383      : 5        |                                        |
     16384 -> 32767      : 1        |                                        |
     32768 -> 65535      : 1        |                                        |
Detaching...
```

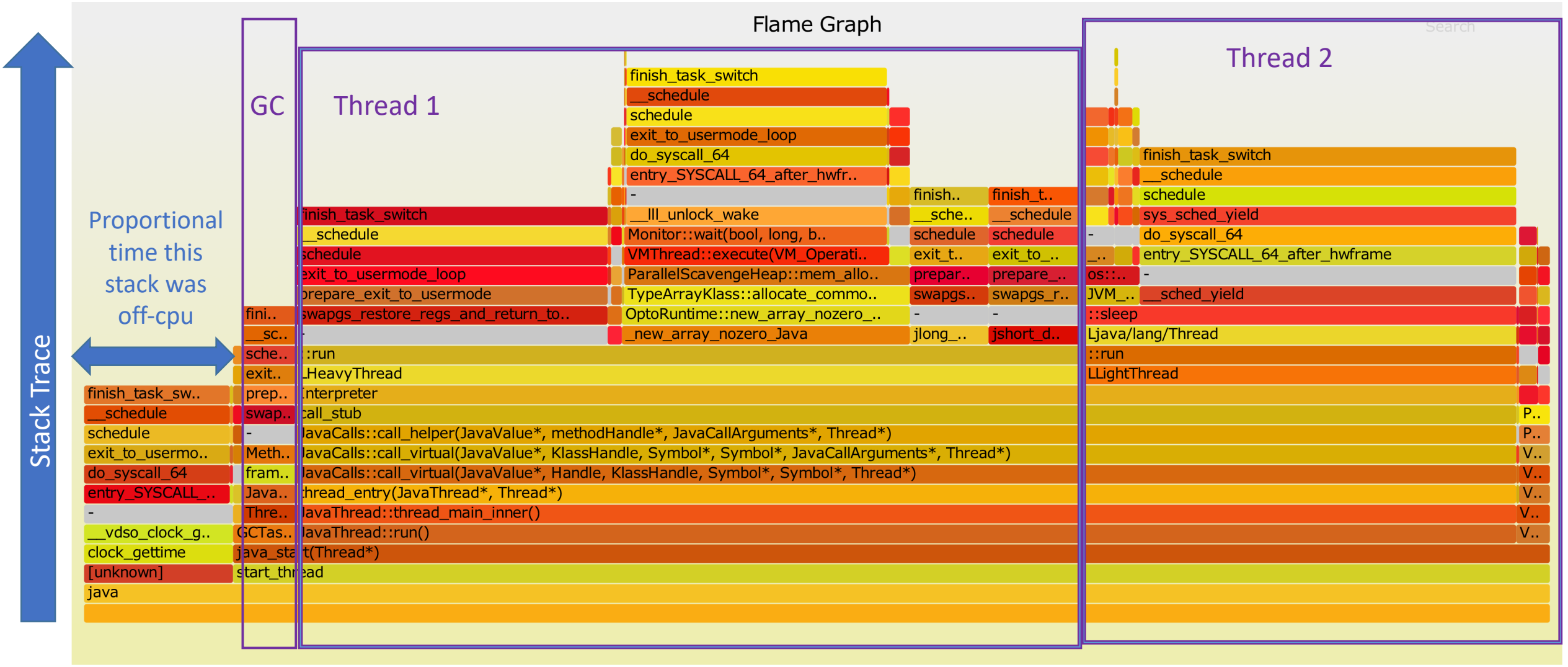
Click to add notes

Example (2/3)
vfs operation stats


KubeCon | CloudNativeCon
Europe 2018

```
# ./vfsstat
TIME         READ/s   WRITE/s CREATE/s    OPEN/s   FSYNC/s
06:46:41:       413         7         0       411         0
06:46:42:        44         3         0         6         0
06:46:43:         2         2         0         0         0
06:46:44:         2         3         0         0         0
06:46:45:       122         3         0        18         0
06:46:46:         2         3         0         0         0
06:46:47:        49         4         0         7         0
```

Example (3/3)
CPU profiles + FlameGraphs



eBPF with containers

kinvolk/**cgnet**

- Attaches eBPF programs to cgroups
- Collects network stats per POD
- Pushes data to Prometheus

Deploying in K8s

- POD spec
  - Privileged pod
  - Share hosts' PID namespace
  - Volume mounts: /lib/modules, /proc
- Expose HTTP endpoints for on-demand collection of matrices

Demo

Kubernetes cluster deployed using Gardener

1. Off-CPU time flamegraph
2. Alerts on shell login
3. Monitoring of PTS device

References

BCC

- https://github.com/iovisor/bcc

FlameGraph

- https://github.com/brendangregg/FlameGraph

Cilium

- https://github.com/cilium/cilium

Weavescope

- https://github.com/weaveworks/tcptracer-bpf

bpfilter

- https://lwn.net/Articles/747551/

Gardener

- https://github.com/gardener/gardener

BCC REST-ified

- https://github.com/ggaurav10/bcc-tools-REST

KubeCon | CloudNativeCon

Europe 2018

Gaurav Gupta

gaurav.gupta07@sap.com

ggaurav10

Questions?