Announcing Mecca

Shachar Shemesh
Chief Court Jester
About the Weka.io product

- “Software only” storage product
- Low latency, high performance
- Written in D
- About 280,000 LoC
  - Not including 114,663 lines in a single auto-generated file.
- Compiled using waf
More About the Code

- Internally called “wekapp”
- Extremely latency sensitive
  - As little GC as possible
  - As few system calls as possible
- Performance sensitive
  - As little copying of data as possible
- Micro-threading (Fibers) based
DPDK and SPDK

• DPDK
  - Intel library for direct DMA to user-space buffer, bypassing the kernel
  - Supported by most high-end NICs
  - Allows getting network data with zero copying

• SPDK
  - s/network/nvme/ in the above description.
Weka Custom Infra

• Why?
  - Phobos relies on GC
  - No standard Fibers scheduler
  - Scheduler must support DPDK polling mode
  - Standard libraries don’t care about performance to our standards

• Weka infra:
  - Busy-polls DPDK and SPDK (optional). Occasionally calls epoll.
  - Containers: Statically allocated, non-GC
  - Time: TSC based, with further performance heuristics.
Technical Debt

• An engineer focused on fixing a bug don’t always focus on all implications of an infrastructure change.
• Decided to repay the debt before it drags us down.
• To prevent further problems: separate into a library.
Distinct Naming Philosophy

- Wekapp – the product repository
- Sir Botty McBotFace – bot for closing integrated bugs
- Teka – tool for deploying from development to AWS
- Deka – the same tool, running inside docker.
Mekka
THE MOMENT I WEKAPP

BEFORE I PUT ON MY MECCAP
What is Mecca?

- A support library
- Containers, libs, the Reactor
- Boost license
- Has a dub package
- Polished against a large code base
Temporary Limitations

- Only for Linux
- Only for x86_64
- API not set in stone
- Primarily aimed at supporting Weka
Where do I get it???

https://github.com/weka-io/mecca
import mecca.log;

@notrace void someFunction(int var)
{
    DEBUG!"Format string %s"(var);
}
Show Me the Mo^H^H Code...
It’s About Time

• What time is it?
• Going to the kernel is expensive
• gettimeofday is better, but still expensive.
• TscTimePoint: directly querying the TSC.
  - Interface change warning: ticks per second will not remain immutable.
Linked List

Head

Next
Prev
Owner (optional)

Next
Prev
Owner (optional)

Next
Prev
Owner (optional)
How much are 6 boys + 3 planets?

```cpp
int someFunction()
{
    int boys = 6;
    int planets = 3;

    return boys + planets;
}
```

```cpp
alias Boys = TypedIdentifier!("Boys", int);
alias Planets = TypedIdentifier!("Planets", int);

int someFunction()
{
    Boys boys = 6;
    Planets planets = 3;

    return boys + planets;
}
```

to!DiskId(diskIdx)
Cascaded Time Queue

1ms

256ms

65.5 seconds

4:40 minutes

49 days

Blue moon events
Mecca’s Fiber Implementation

- D’s Fiber model requires jumping into and out of the fiber.
  - Two register set switches per context switch.
- Mecca switch directly to next fiber’s context.
  - Save only those registers that are not clobbered by the ABI.
  - Does not save the floating point registers.
Generic Reactor Flow

```c
ssize_t read(int fd, params) {
    ssize_t ret;
    while(
        (ret=read(fd, params))<0 &&
        (errno==EAGAIN || errno==EWOULDBLOCK) )
    {
        registerForRead(fd);
        yield();
        unregisterForRead(fd);  // EPOLLONESHOT
    }
    return ret;
}
```
Edge Trigger IO Switching

- Epoll has an “edge trigger” mode.
- Considered almost useless
- Actually matches the fibers’ working mode like a glove.
Generic Reactor Flow

```c
ssize_t read(int fd, params) {
    ssize_t ret;
    while(
        (ret=read(fd, params))<0 &&
        (errno==EAGAIN || errno==EWOULDBLOCK) )
    {
        registerForRead(fd);  // EPOLLET
        yield();
        unregisterForRead(fd); // EPOLLET
    }

    return ret;
}
```