Scaling the Boot Barrier: Identifying and Eliminating Contention in OpenStack

gridcentric

Peter Feiner
peter@gridcentric.com

Applications as VMs

Applications as VMs

- Applications deployed in virtual machines
  - Carve up big hosts
  - Makes application capacity granular

Applications as VMs

- Applications deployed in virtual machines
  - Carve up big hosts
  - Makes application capacity granular
- Increase capacity by creating more VMs
  - Create more VMs as load approaches capacity
  - When should you create more?

When to Create More

When to Create More

- As late as possible
  - Avoid over provisioning

When to Create More

- As late as possible
  - Avoid over provisioning
- As soon as necessary
  - Anticipate when load will surpass capacity
  - Factor in time it takes for new VM start serving
    - How can we optimize this (i.e., make it low)?

Time to Start Serving

VM Creation Time + Guest preparation time

- Time for OS to boot and app to start serving
- Lean OS & stateless app can serve in < 10s
- Fat OS & big app ready instantly with **live images**

Time to Start Serving

VM Creation Time + Guest preparation time

- Time from `nova boot` to `ACTIVE`

- Time for OS to boot and app to start serving
- Lean OS & stateless app can serve in < 10s
- Fat OS & big app ready instantly with **live images**

Time to Start Serving

VM Creation Time + Guest preparation time

- Time from `nova boot` to `ACTIVE`

- Time for OS to boot and app to start serving

Time to Start Serving

VM Creation Time + Guest preparation time

- Time from `nova boot` to `ACTIVE`

- Time for OS to boot and app to start serving
- Lean OS & stateless app can serve in < 10s

Time to Start Serving

VM Creation Time + Guest preparation time

- Time from `nova boot` to `ACTIVE`

- Time for OS to boot and app to start serving
- Lean OS & stateless app can serve in < 10s
- Fat OS & big app ready instantly with **live images**

Time to Start Serving

VM Creation Time +

- Time from `nova boot` to `ACTIVE`

Guest preparation time

- Time for OS to boot and app to start serving
- Lean OS & stateless app can serve in < 10s
- Fat OS & big app ready instantly with **live images**

gridcentric

Time to Start Serving

VM Creation Time + Guest preparation time

VM Creation Time

- Time from `nova boot` to `ACTIVE`
- Can take a long time

Guest preparation time

- Time for OS to boot and app to start serving
- Lean OS & stateless app can serve in < 10s
- Fat OS & big app ready instantly with **live images**

gridcentric

Time to Start Serving

VM Creation Time + Guest preparation time

- Time from `nova boot` to `ACTIVE`
- Can take a long time
- Let’s do an experiment ...

- Time for OS to boot and app to start serving
- Lean OS & stateless app can serve in < 10s
- Fat OS & big app ready instantly with **live images**

gridcentric

Experimental Setup

Experimental Setup

- Create VMs in parallel
  - Make *N* creation requests in parallel
  - Measure time from API request to ACTIVE

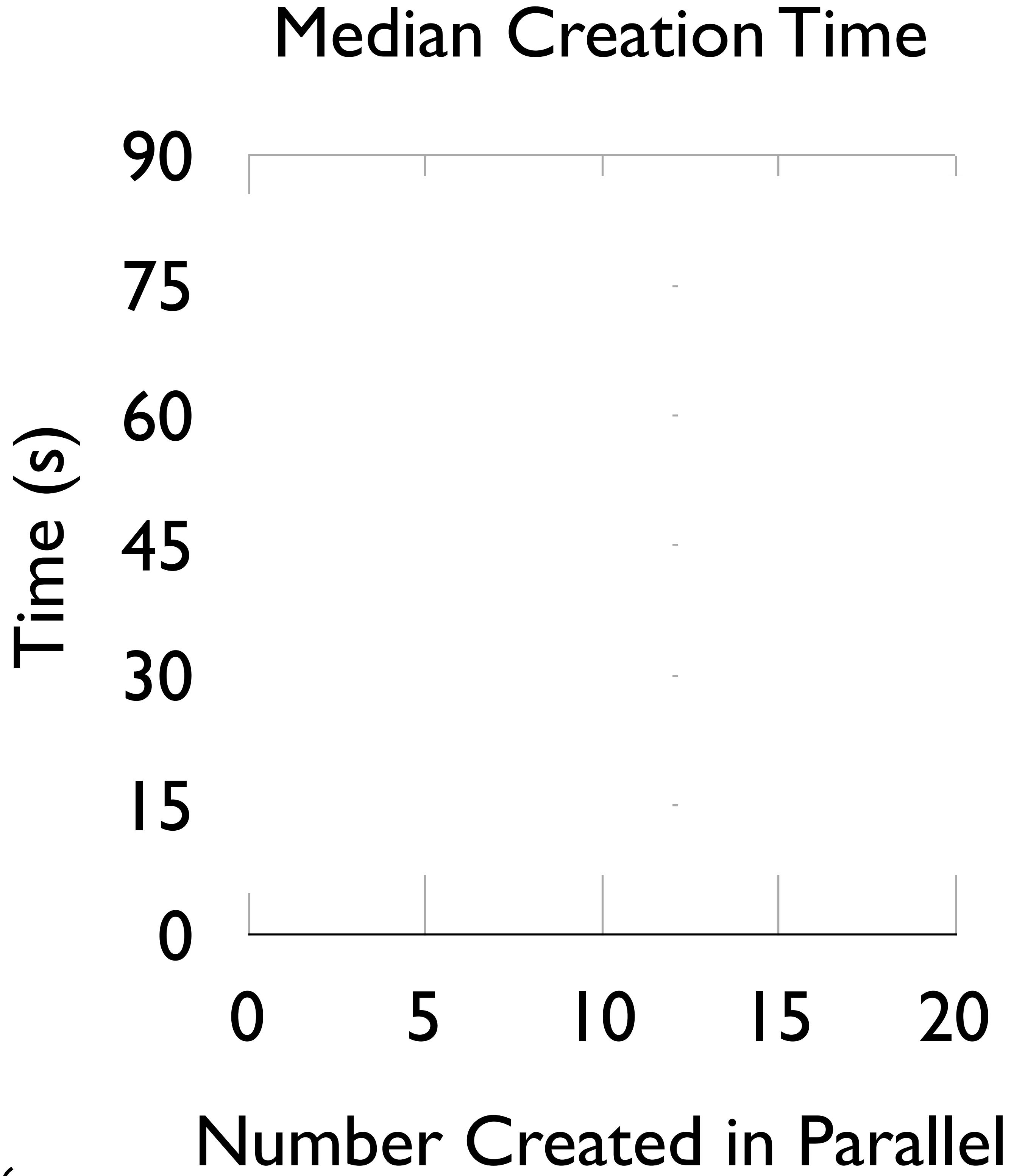
Experimental Setup

- Create VMs in parallel
  - Make *N* creation requests in parallel
  - Measure time from API request to ACTIVE
- OpenStack Grizzly
  - Compute: Libvirt + KVM
  - Networking: Quantum + Open vSwitch
  - Storage: qcow2

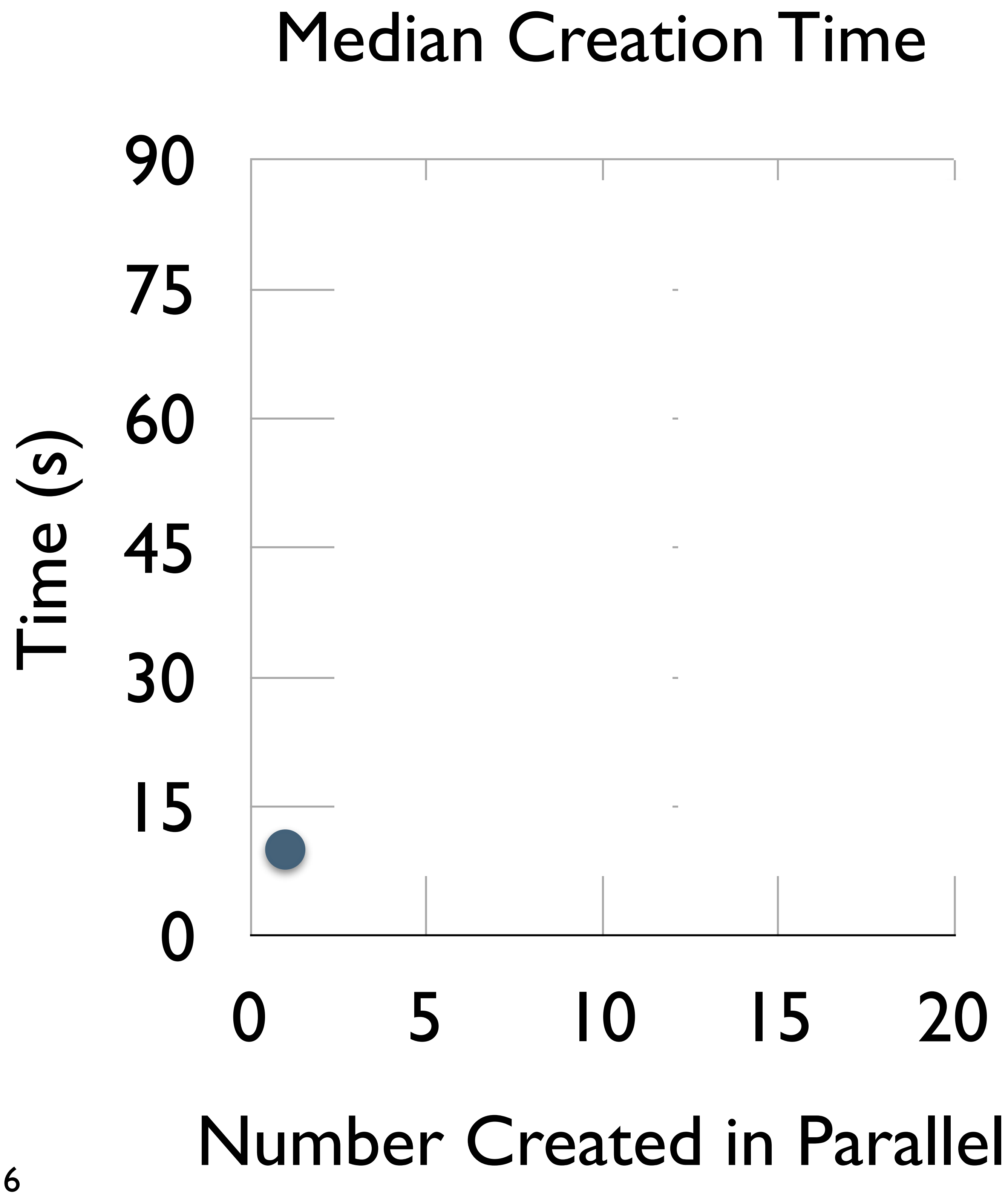
Experimental Setup

- Create VMs in parallel
  - Make *N* creation requests in parallel
  - Measure time from API request to ACTIVE
- OpenStack Grizzly
  - Compute: Libvirt + KVM
  - Networking: Quantum + Open vSwitch
  - Storage: qcow2
- 96 GB RAM, 12 cores x 2 HT/core, SSD

VM Creation Time

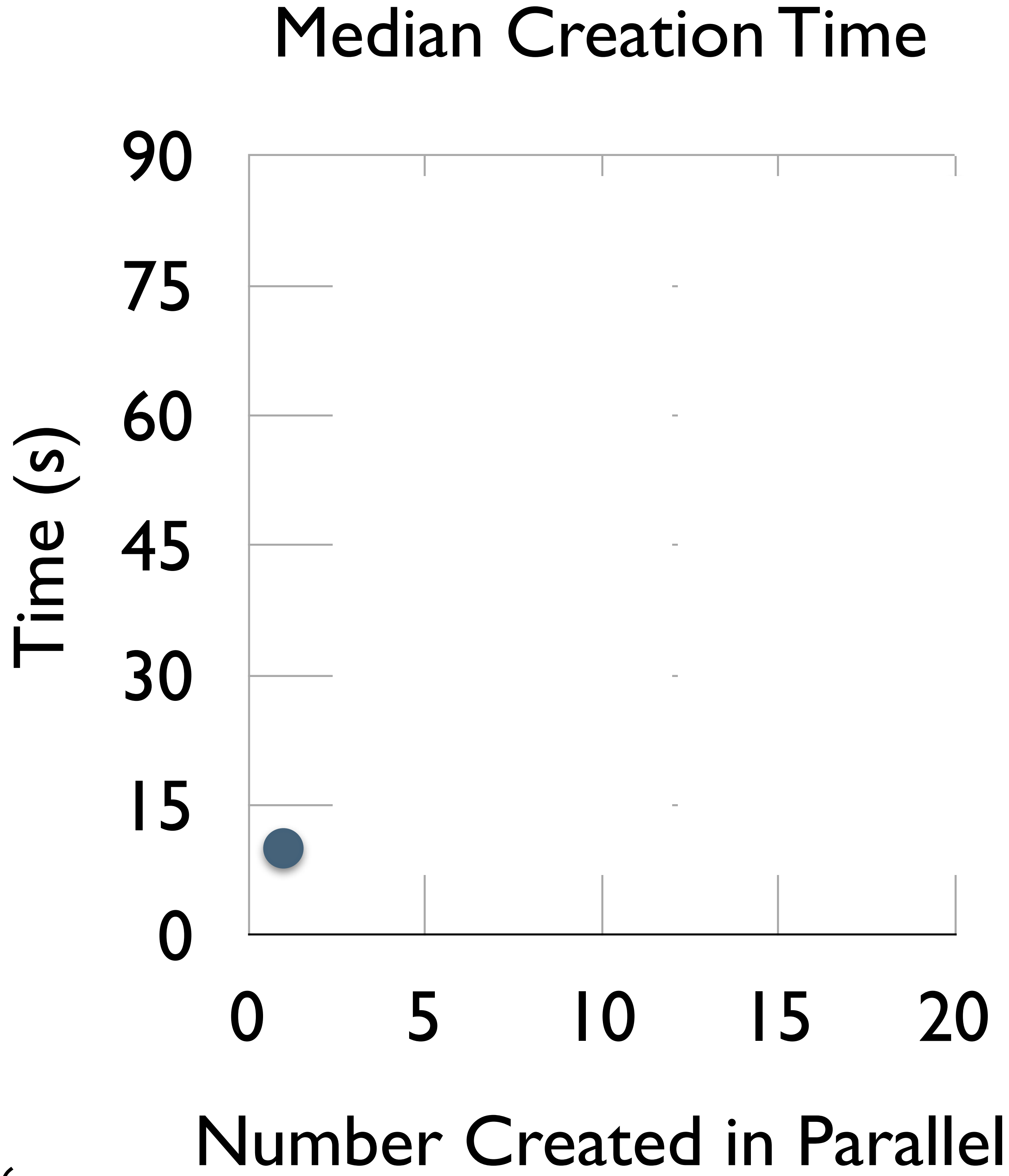


VM Creation Time



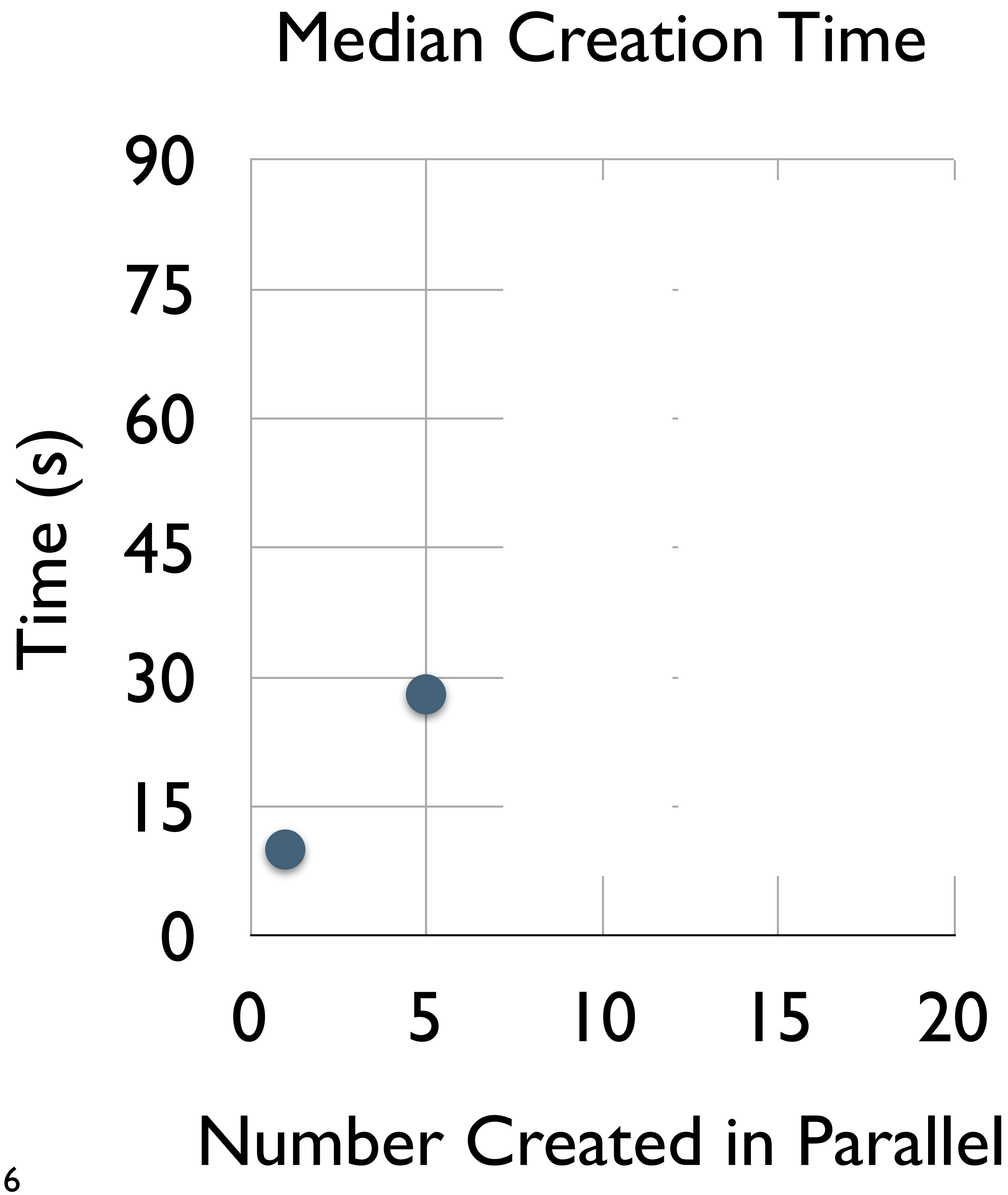
VM Creation Time

- Single VM is fast ~10s



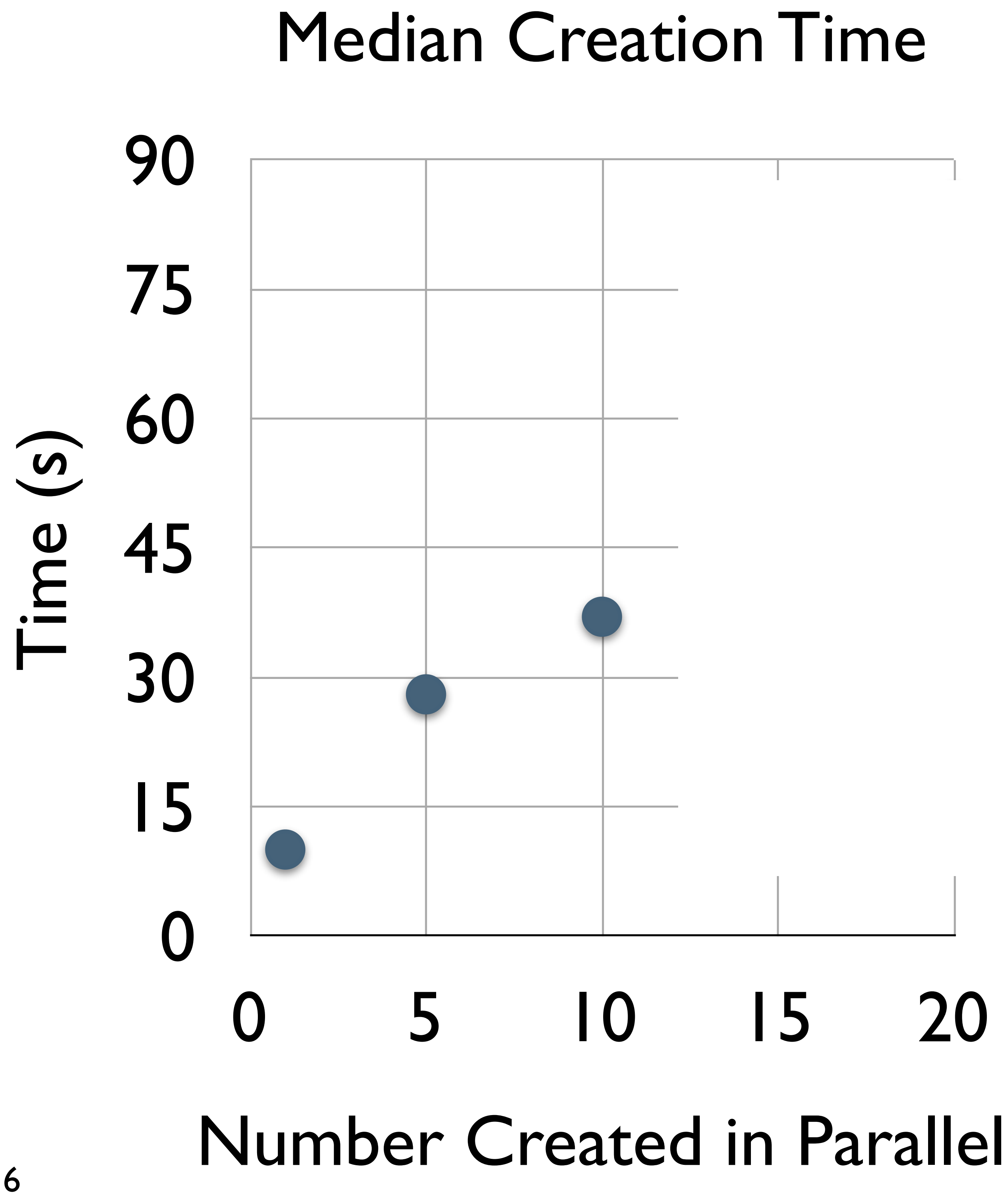
VM Creation Time

- Single VM is fast ~10s



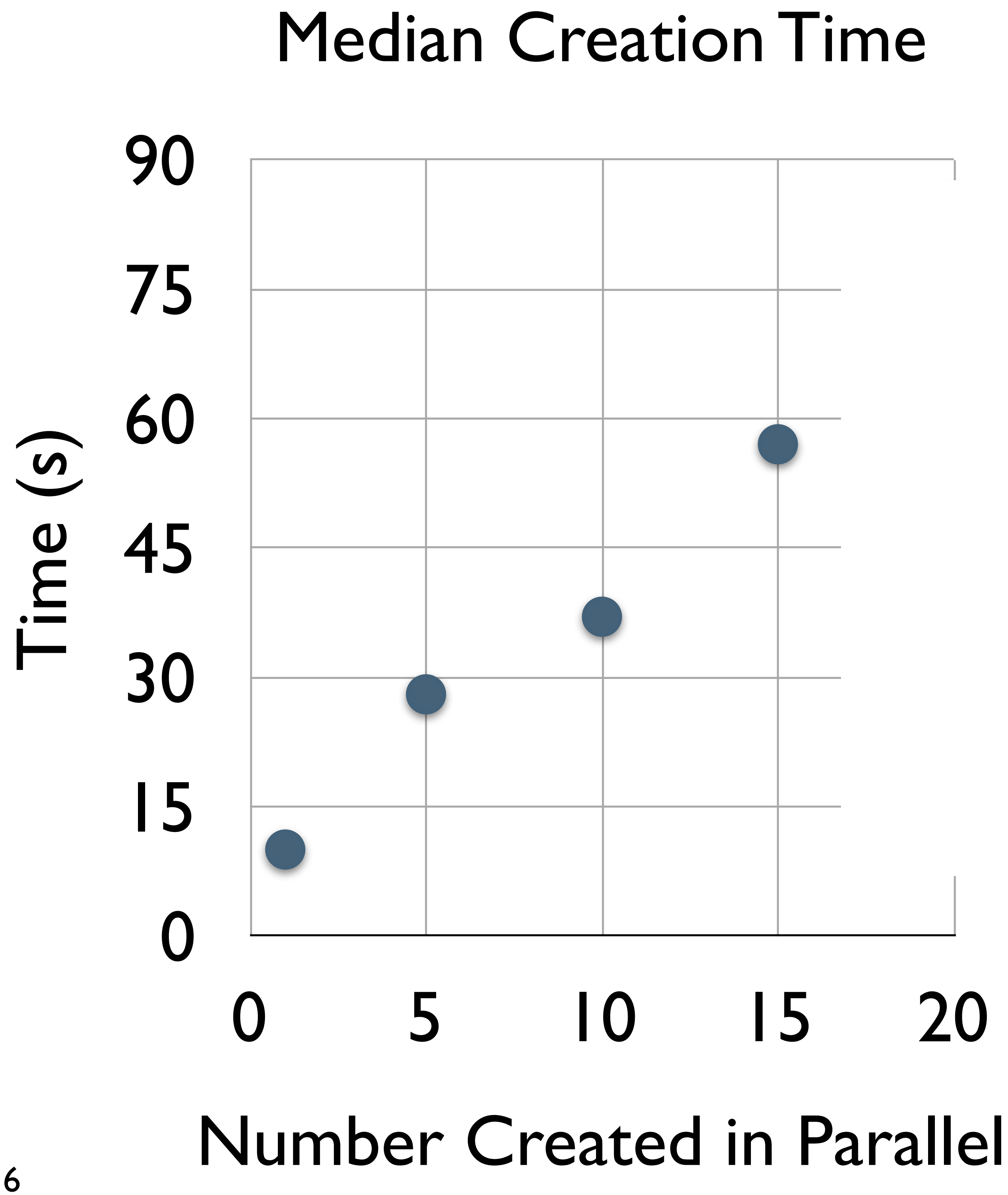
VM Creation Time

- Single VM is fast ~10s



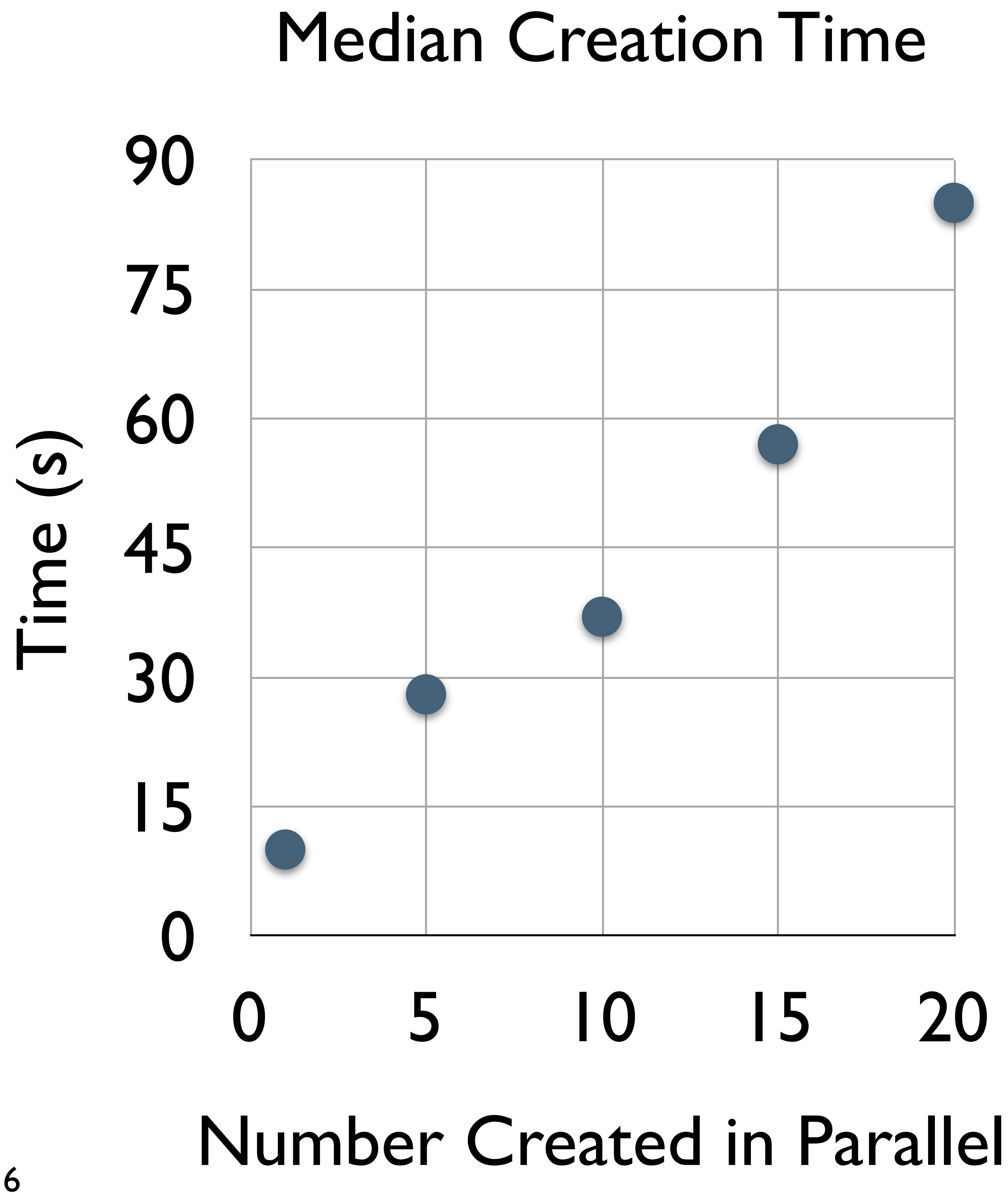
VM Creation Time

- Single VM is fast ~10s



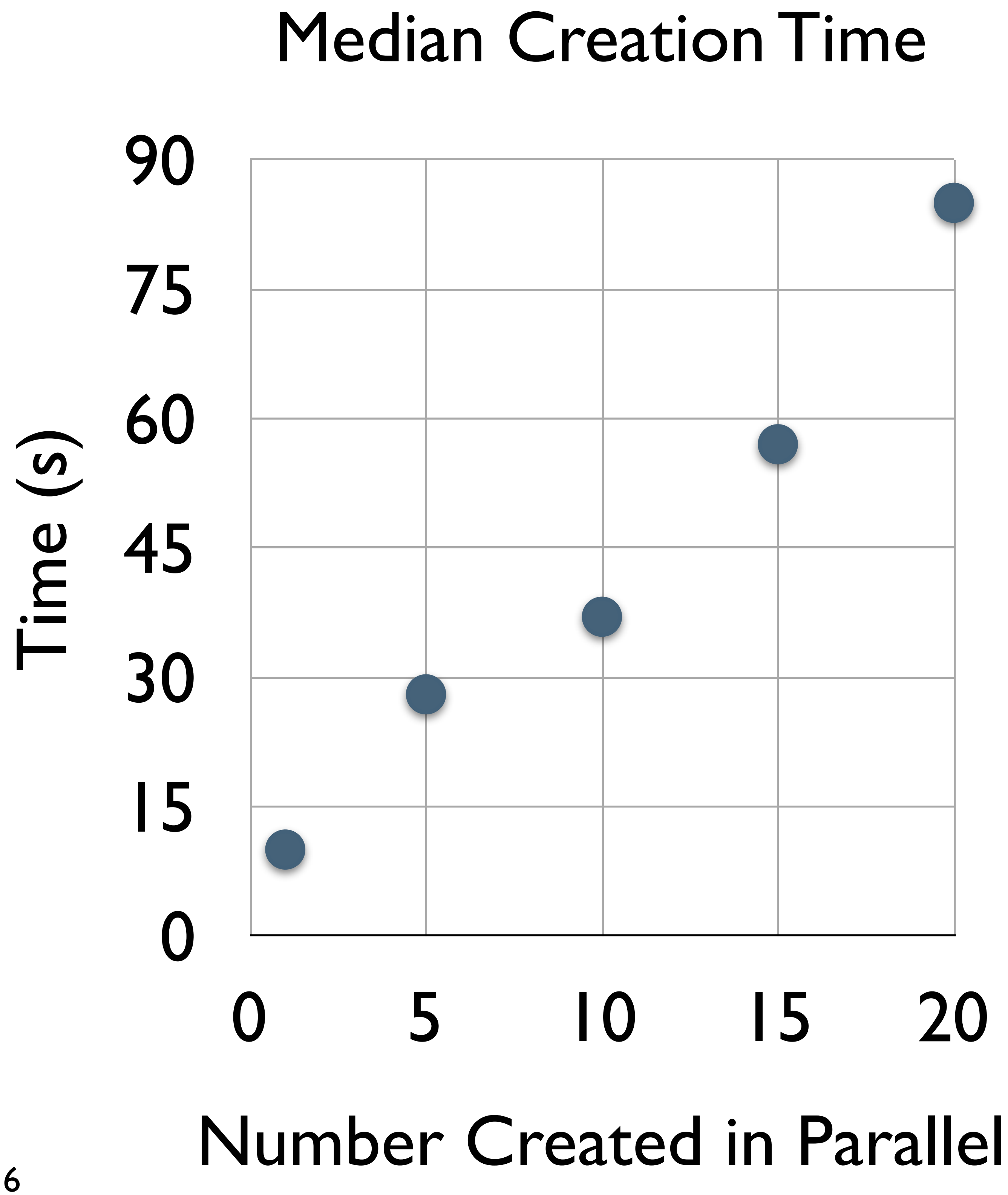
VM Creation Time

- Single VM is fast ~10s



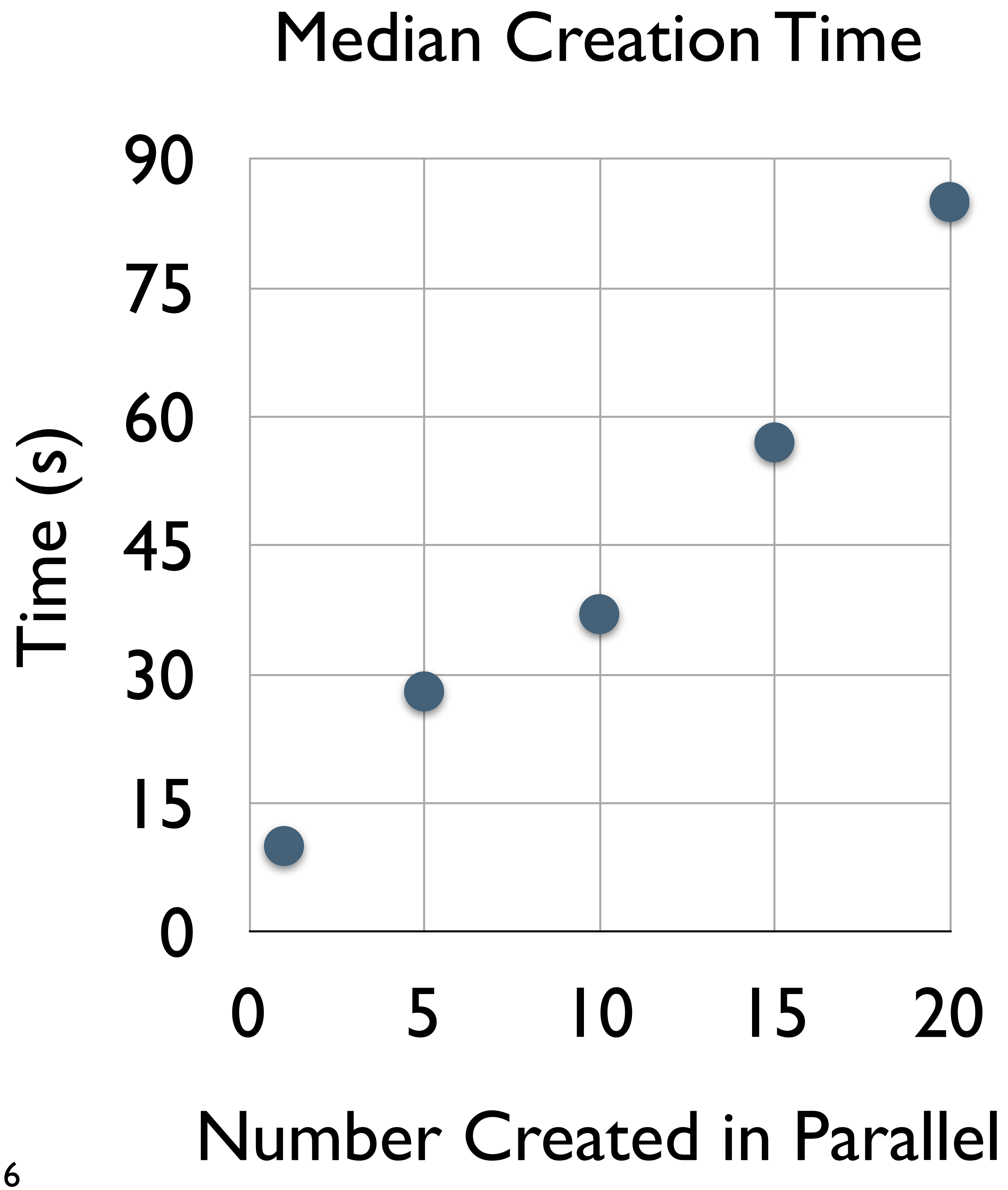
VM Creation Time

- Single VM is fast ~10s
- Many VMs can be slow
  - Creation time increases linearly with *N*
  - Must be some bottlenecks



VM Creation Time

- Single VM is fast ~10s
- Many VMs can be slow
  - Creation time increases linearly with *N*
  - Must be some bottlenecks
- Looks worse without quantum
  - 10s longer when *N*=20



Possible Bottlenecks

Possible Bottlenecks

- Hardware
  - CPUs pegged? RAM all used? Disk busy?

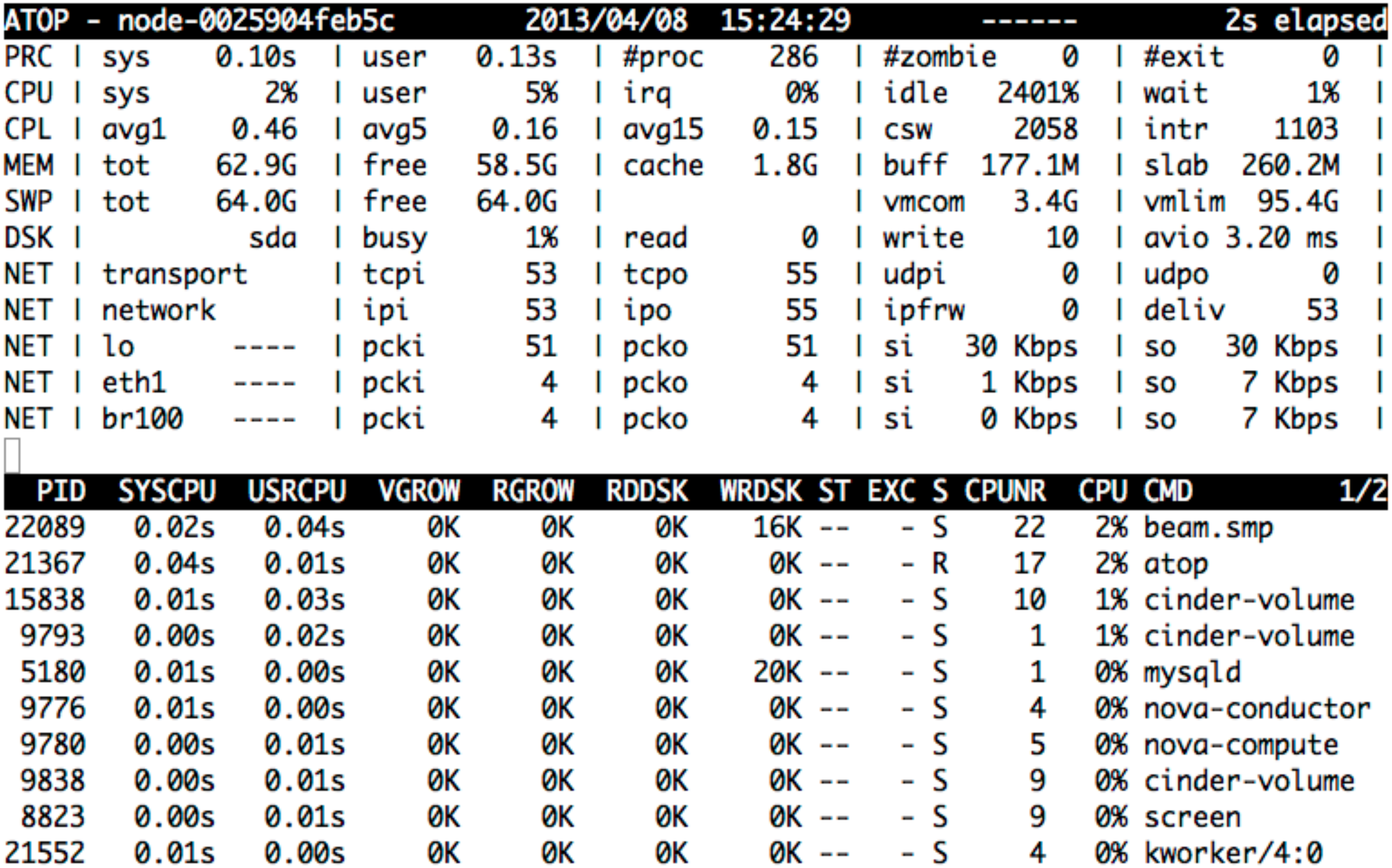
Possible Bottlenecks

- Hardware
  - CPUs pegged? RAM all used? Disk busy?
- Software
  - Locks held for a long time?

Possible Bottlenecks

- Hardware
  - CPUs pegged? RAM all used? Disk busy?
- Software
  - Locks held for a long time?
- Hardware easy to check with `atop`
  - Let's look at `atop` first

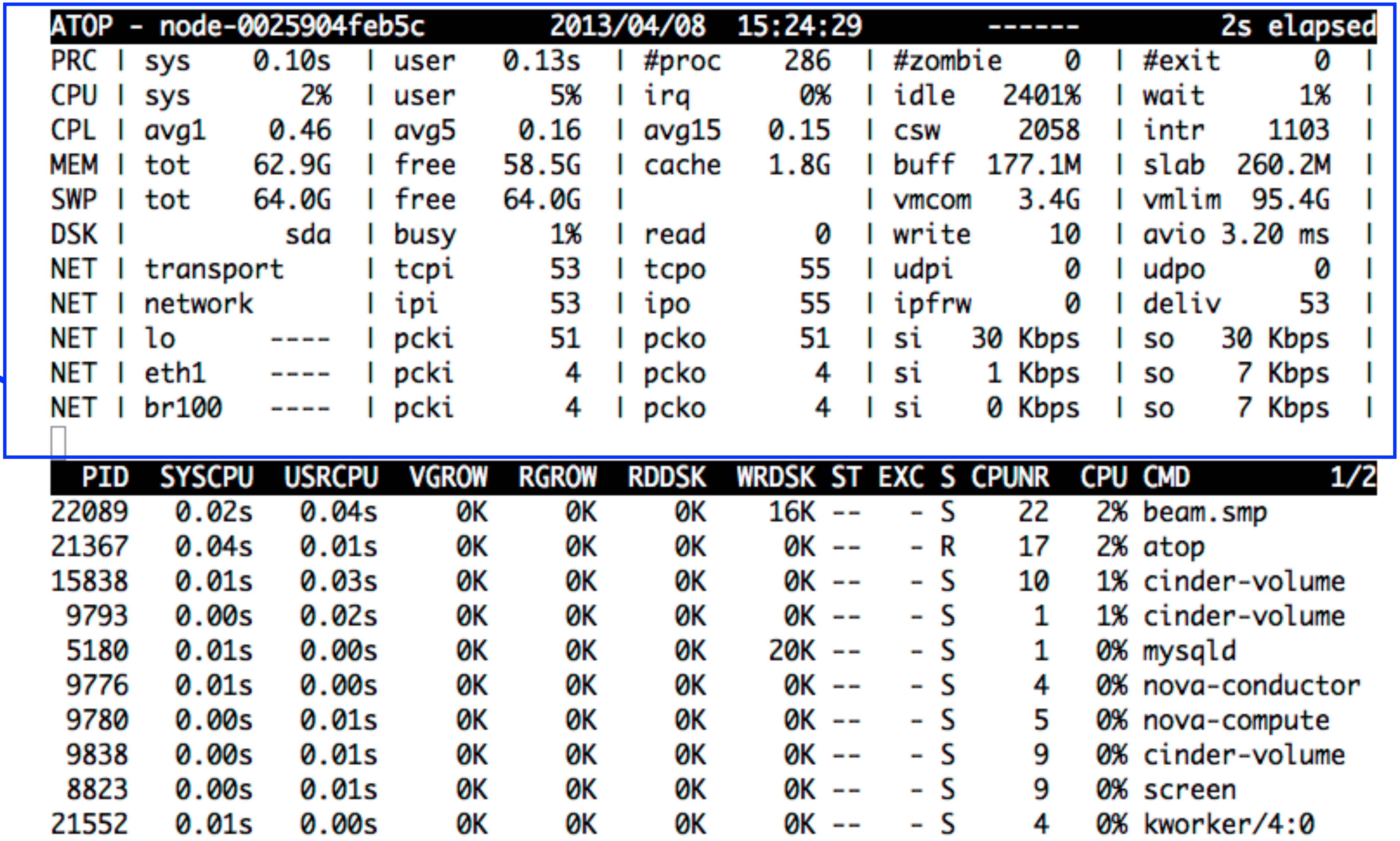
atop



```
ATOP - node-0025904feb5c        2013/04/08  15:24:29        ------         2s elapsed
PRC | sys    0.10s  | user   0.13s  | #proc    286  | #zombie    0  | #exit      0  |
CPU | sys       2%  | user      5%  | irq       0%  | idle   2401%  | wait      1%  |
CPL | avg1    0.46  | avg5    0.16  | avg15   0.15  | csw     2058  | intr    1103  |
MEM | tot    62.9G  | free   58.5G  | cache   1.8G  | buff  177.1M  | slab  260.2M  |
SWP | tot    64.0G  | free   64.0G  |               | vmcom   3.4G  | vmlim  95.4G  |
DSK |          sda  | busy      1%  | read       0  | write     10  | avio 3.20 ms  |
NET | transport     | tcpi      53  | tcpo      55  | udpi       0  | udpo       0  |
NET | network       | ipi       53  | ipo       55  | ipfrw      0  | deliv     53  |
NET | lo      ----  | pcki      51  | pcko      51  | si   30 Kbps  | so   30 Kbps  |
NET | eth1    ----  | pcki       4  | pcko       4  | si    1 Kbps  | so    7 Kbps  |
NET | br100   ----  | pcki       4  | pcko       4  | si    0 Kbps  | so    7 Kbps  |

  PID  SYSCPU  USRCPU  VGROW  RGROW  RDDSK  WRDSK ST EXC S CPUNR  CPU CMD         1/2
22089   0.02s   0.04s     0K     0K     0K    16K --   - S    22   2% beam.smp
21367   0.04s   0.01s     0K     0K     0K     0K --   - R    17   2% atop
15838   0.01s   0.03s     0K     0K     0K     0K --   - S    10   1% cinder-volume
 9793   0.00s   0.02s     0K     0K     0K     0K --   - S     1   1% cinder-volume
 5180   0.01s   0.00s     0K     0K     0K    20K --   - S     1   0% mysqld
 9776   0.01s   0.00s     0K     0K     0K     0K --   - S     4   0% nova-conductor
 9780   0.00s   0.01s     0K     0K     0K     0K --   - S     5   0% nova-compute
 9838   0.00s   0.01s     0K     0K     0K     0K --   - S     9   0% cinder-volume
 8823   0.00s   0.01s     0K     0K     0K     0K --   - S     9   0% screen
21552   0.01s   0.00s     0K     0K     0K     0K --   - S     4   0% kworker/4:0
```

atop



System Wide

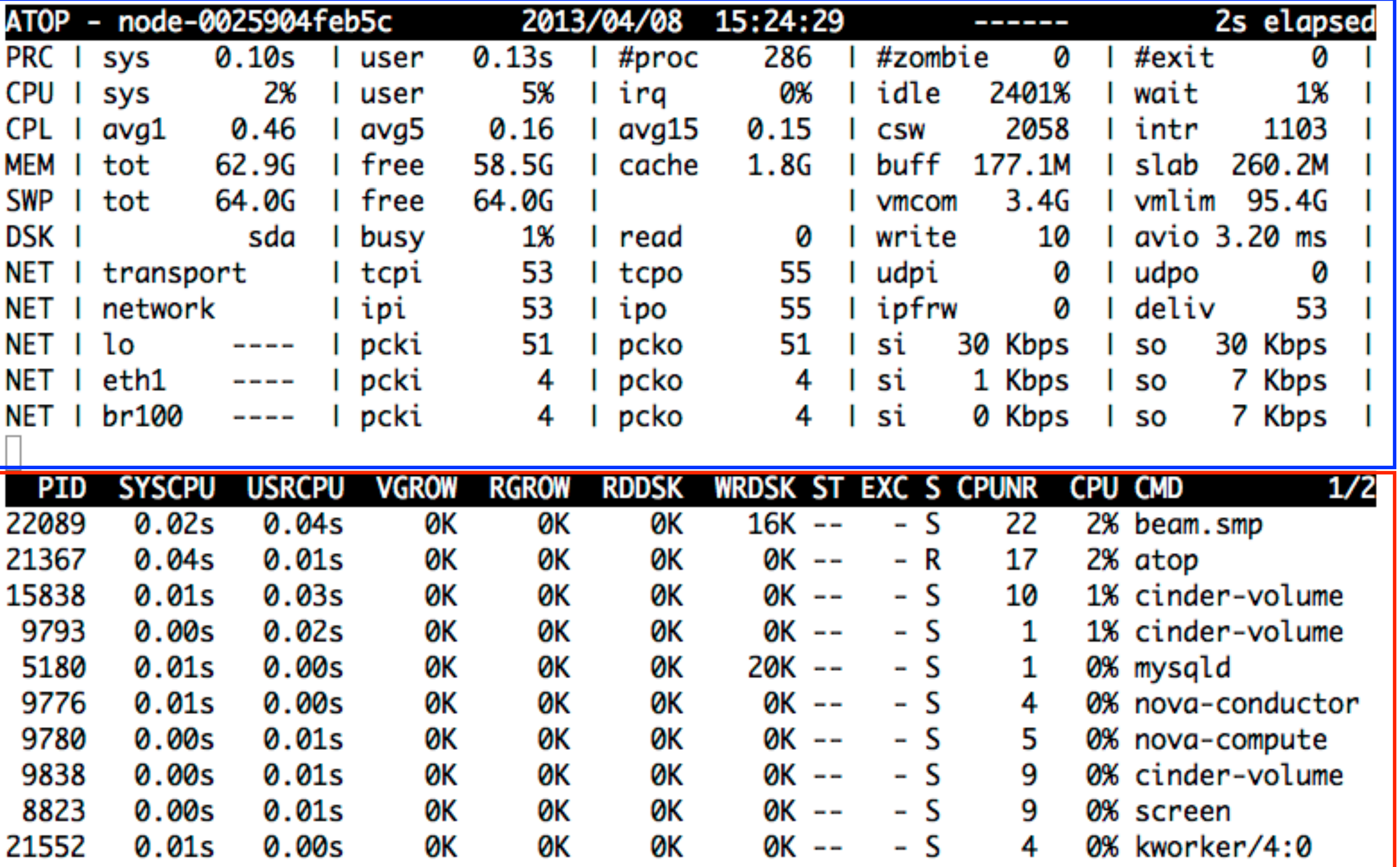
```
ATOP - node-0025904feb5c        2013/04/08  15:24:29         ------          2s elapsed
PRC | sys     0.10s  | user    0.13s  | #proc     286  | #zombie     0  | #exit      0  |
CPU | sys        2%  | user       5%  | irq        0%  | idle    2401%  | wait      1%  |
CPL | avg1     0.46  | avg5     0.16  | avg15    0.15  | csw      2058  | intr    1103  |
MEM | tot     62.9G  | free    58.5G  | cache    1.8G  | buff   177.1M  | slab  260.2M  |
SWP | tot     64.0G  | free    64.0G  |                | vmcom    3.4G  | vmlim  95.4G  |
DSK |           sda  | busy       1%  | read        0  | write      10  | avio 3.20 ms  |
NET | transport      | tcpi       53  | tcpo       55  | udpi        0  | udpo       0  |
NET | network        | ipi        53  | ipo        55  | ipfrw       0  | deliv     53  |
NET | lo       ----  | pcki       51  | pcko       51  | si    30 Kbps  | so   30 Kbps  |
NET | eth1     ----  | pcki        4  | pcko        4  | si     1 Kbps  | so    7 Kbps  |
NET | br100    ----  | pcki        4  | pcko        4  | si     0 Kbps  | so    7 Kbps  |

  PID  SYSCPU  USRCPU  VGROW  RGROW  RDDSK  WRDSK ST EXC S CPUNR  CPU CMD            1/2
22089   0.02s   0.04s     0K     0K     0K    16K --   - S    22   2% beam.smp
21367   0.04s   0.01s     0K     0K     0K     0K --   - R    17   2% atop
15838   0.01s   0.03s     0K     0K     0K     0K --   - S    10   1% cinder-volume
 9793   0.00s   0.02s     0K     0K     0K     0K --   - S     1   1% cinder-volume
 5180   0.01s   0.00s     0K     0K     0K    20K --   - S     1   0% mysqld
 9776   0.01s   0.00s     0K     0K     0K     0K --   - S     4   0% nova-conductor
 9780   0.00s   0.01s     0K     0K     0K     0K --   - S     5   0% nova-compute
 9838   0.00s   0.01s     0K     0K     0K     0K --   - S     9   0% cinder-volume
 8823   0.00s   0.01s     0K     0K     0K     0K --   - S     9   0% screen
21552   0.01s   0.00s     0K     0K     0K     0K --   - S     4   0% kworker/4:0
```

atop

System Wide

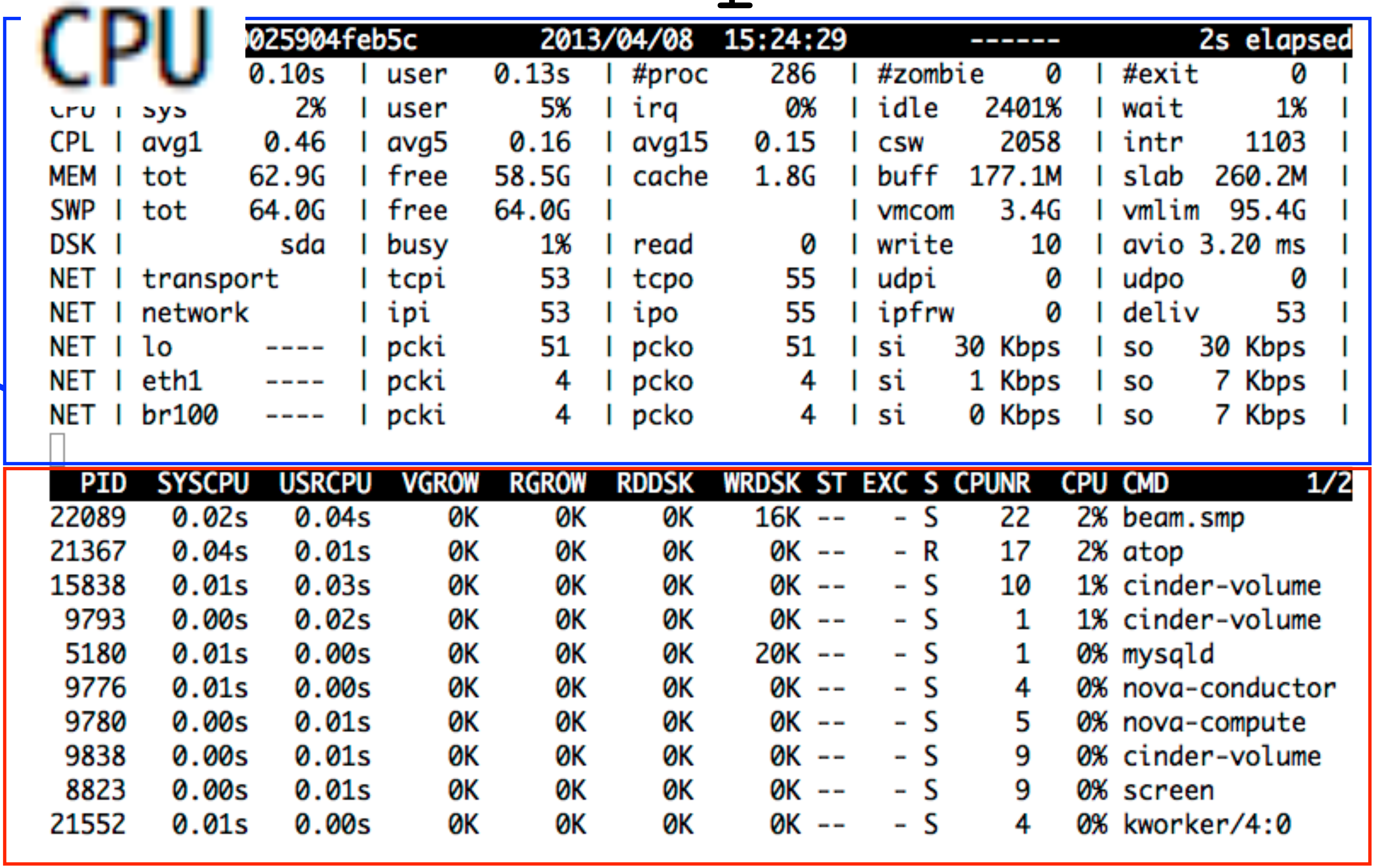
Per Process



```
ATOP - node-0025904feb5c       2013/04/08  15:24:29        -------          2s elapsed
PRC | sys    0.10s | user   0.13s | #proc    286 | #zombie    0 | #exit      0 |
CPU | sys       2% | user      5% | irq       0% | idle   2401% | wait      1% |
CPL | avg1    0.46 | avg5    0.16 | avg15   0.15 | csw     2058 | intr    1103 |
MEM | tot    62.9G | free   58.5G | cache   1.8G | buff  177.1M | slab  260.2M |
SWP | tot    64.0G | free   64.0G |              | vmcom   3.4G | vmlim  95.4G |
DSK |          sda | busy      1% | read       0 | write     10 | avio 3.20 ms |
NET | transport    | tcpi      53 | tcpo      55 | udpi       0 | udpo       0 |
NET | network      | ipi       53 | ipo       55 | ipfrw      0 | deliv     53 |
NET | lo      ---- | pcki      51 | pcko      51 | si   30 Kbps | so   30 Kbps |
NET | eth1    ---- | pcki       4 | pcko       4 | si    1 Kbps | so    7 Kbps |
NET | br100   ---- | pcki       4 | pcko       4 | si    0 Kbps | so    7 Kbps |

  PID  SYSCPU  USRCPU  VGROW  RGROW  RDDSK  WRDSK ST EXC S CPUNR  CPU CMD        1/2
22089   0.02s   0.04s     0K     0K     0K    16K --   - S    22   2% beam.smp
21367   0.04s   0.01s     0K     0K     0K     0K --   - R    17   2% atop
15838   0.01s   0.03s     0K     0K     0K     0K --   - S    10   1% cinder-volume
 9793   0.00s   0.02s     0K     0K     0K     0K --   - S     1   1% cinder-volume
 5180   0.01s   0.00s     0K     0K     0K    20K --   - S     1   0% mysqld
 9776   0.01s   0.00s     0K     0K     0K     0K --   - S     4   0% nova-conductor
 9780   0.00s   0.01s     0K     0K     0K     0K --   - S     5   0% nova-compute
 9838   0.00s   0.01s     0K     0K     0K     0K --   - S     9   0% cinder-volume
 8823   0.00s   0.01s     0K     0K     0K     0K --   - S     9   0% screen
21552   0.01s   0.00s     0K     0K     0K     0K --   - S     4   0% kworker/4:0
```

atop



System Wide

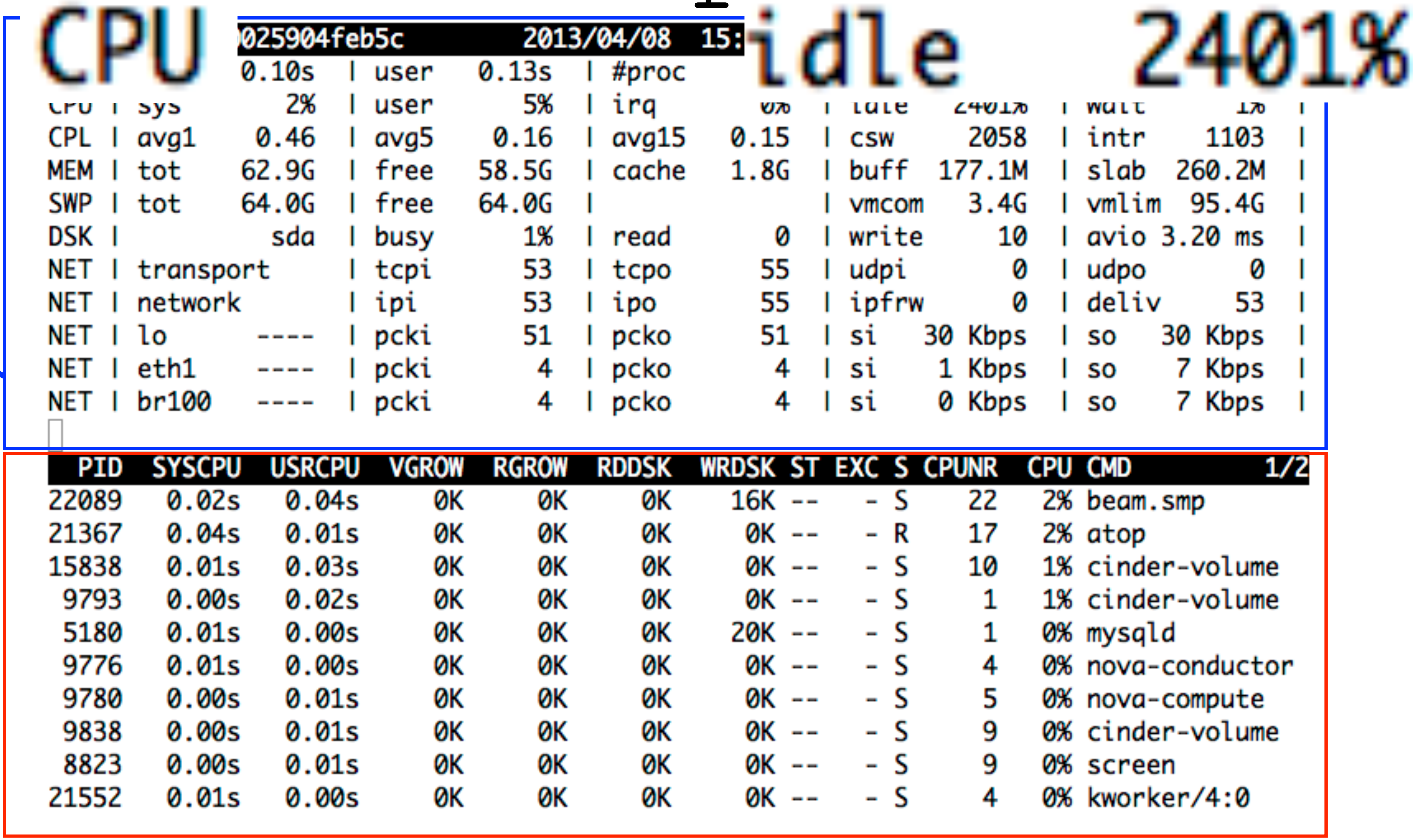
CPU

```
 025904feb5c          2013/04/08  15:24:29          -------            2s elapsed
               0.10s | user   0.13s | #proc    286 | #zombie     0 | #exit      0 |
CPU | sys        2% | user      5% | irq       0% | idle   2401% | wait      1% |
CPL | avg1     0.46 | avg5    0.16 | avg15   0.15 | csw     2058 | intr    1103 |
MEM | tot     62.9G | free   58.5G | cache   1.8G | buff  177.1M | slab  260.2M |
SWP | tot     64.0G | free   64.0G |              | vmcom   3.4G | vmlim  95.4G |
DSK |           sda | busy      1% | read       0 | write     10 | avio 3.20 ms |
NET | transport     | tcpi      53 | tcpo      55 | udpi       0 | udpo       0 |
NET | network       | ipi       53 | ipo       55 | ipfrw      0 | deliv     53 |
NET | lo       ---- | pcki      51 | pcko      51 | si   30 Kbps | so   30 Kbps |
NET | eth1     ---- | pcki       4 | pcko       4 | si    1 Kbps | so    7 Kbps |
NET | br100    ---- | pcki       4 | pcko       4 | si    0 Kbps | so    7 Kbps |
```

Per Process

```
  PID  SYSCPU  USRCPU  VGROW  RGROW  RDDSK  WRDSK ST EXC S CPUNR  CPU CMD          1/2
22089   0.02s   0.04s     0K     0K     0K    16K --   - S    22   2% beam.smp
21367   0.04s   0.01s     0K     0K     0K     0K --   - R    17   2% atop
15838   0.01s   0.03s     0K     0K     0K     0K --   - S    10   1% cinder-volume
 9793   0.00s   0.02s     0K     0K     0K     0K --   - S     1   1% cinder-volume
 5180   0.01s   0.00s     0K     0K     0K    20K --   - S     1   0% mysqld
 9776   0.01s   0.00s     0K     0K     0K     0K --   - S     4   0% nova-conductor
 9780   0.00s   0.01s     0K     0K     0K     0K --   - S     5   0% nova-compute
 9838   0.00s   0.01s     0K     0K     0K     0K --   - S     9   0% cinder-volume
 8823   0.00s   0.01s     0K     0K     0K     0K --   - S     9   0% screen
21552   0.01s   0.00s     0K     0K     0K     0K --   - S     4   0% kworker/4:0
```

atop



System Wide

```
CPU                                              idle        2401%
    025904feb5c          2013/04/08   15:
    0.10s  | user    0.13s  | #proc
cpu | sys      2%  | user      5%  | irq       0%  | idle   2401%  | wait      1%  |
CPL | avg1   0.46  | avg5    0.16  | avg15   0.15  | csw     2058  | intr    1103  |
MEM | tot   62.9G  | free   58.5G  | cache   1.8G  | buff  177.1M  | slab  260.2M  |
SWP | tot   64.0G  | free   64.0G  |               | vmcom   3.4G  | vmlim  95.4G  |
DSK |         sda  | busy      1%  | read       0  | write     10  | avio 3.20 ms  |
NET | transport    | tcpi      53  | tcpo      55  | udpi       0  | udpo       0  |
NET | network      | ipi       53  | ipo       55  | ipfrw      0  | deliv     53  |
NET | lo      ---- | pcki      51  | pcko      51  | si   30 Kbps  | so   30 Kbps  |
NET | eth1    ---- | pcki       4  | pcko       4  | si    1 Kbps  | so    7 Kbps  |
NET | br100   ---- | pcki       4  | pcko       4  | si    0 Kbps  | so    7 Kbps  |
```

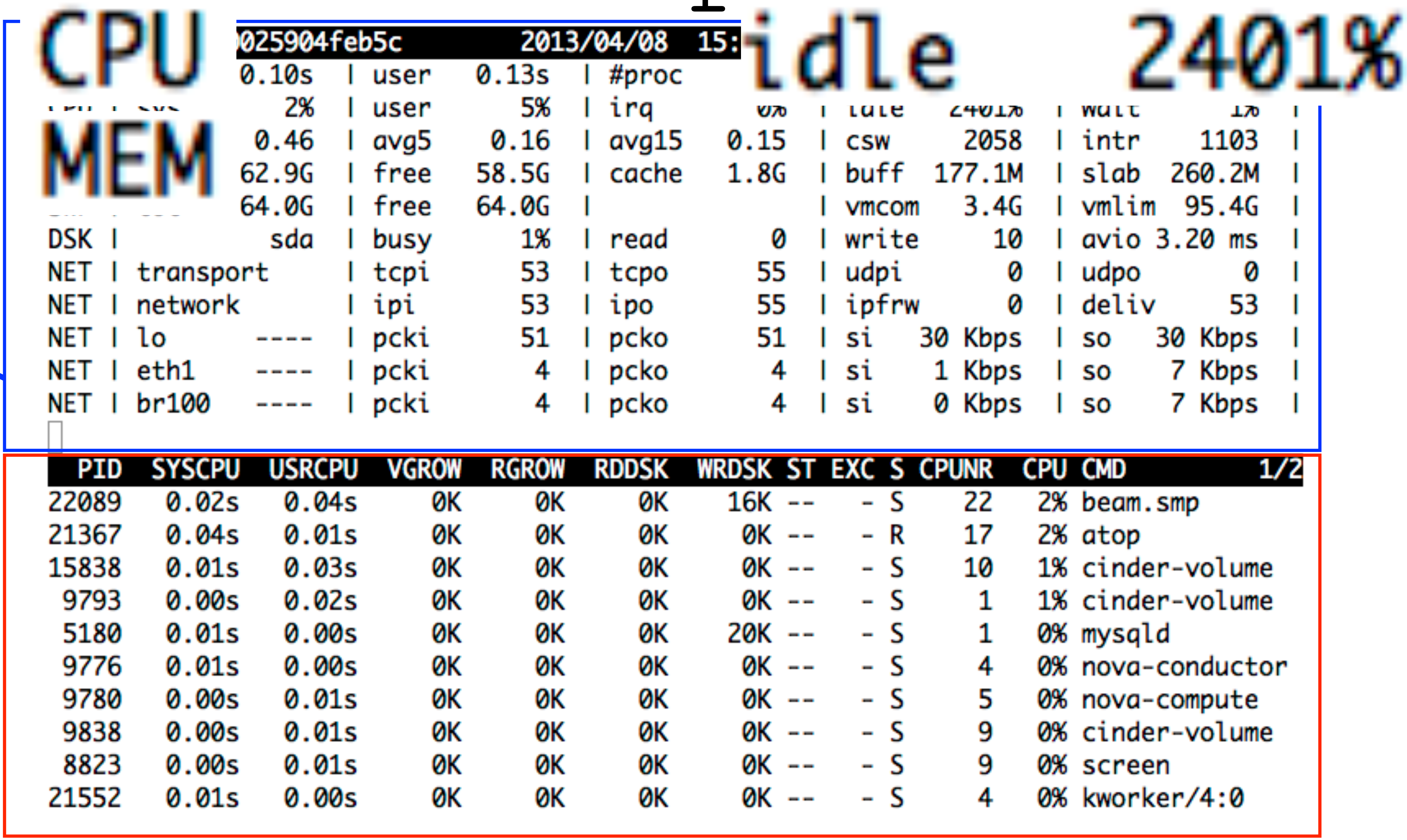
Per Process

```
  PID  SYSCPU  USRCPU  VGROW  RGROW  RDDSK  WRDSK ST EXC S CPUNR  CPU CMD          1/2
22089   0.02s   0.04s     0K     0K     0K    16K --   - S    22   2% beam.smp
21367   0.04s   0.01s     0K     0K     0K     0K --   - R    17   2% atop
15838   0.01s   0.03s     0K     0K     0K     0K --   - S    10   1% cinder-volume
 9793   0.00s   0.02s     0K     0K     0K     0K --   - S     1   1% cinder-volume
 5180   0.01s   0.00s     0K     0K     0K    20K --   - S     1   0% mysqld
 9776   0.01s   0.00s     0K     0K     0K     0K --   - S     4   0% nova-conductor
 9780   0.00s   0.01s     0K     0K     0K     0K --   - S     5   0% nova-compute
 9838   0.00s   0.01s     0K     0K     0K     0K --   - S     9   0% cinder-volume
 8823   0.00s   0.01s     0K     0K     0K     0K --   - S     9   0% screen
21552   0.01s   0.00s     0K     0K     0K     0K --   - S     4   0% kworker/4:0
```

atop

System Wide

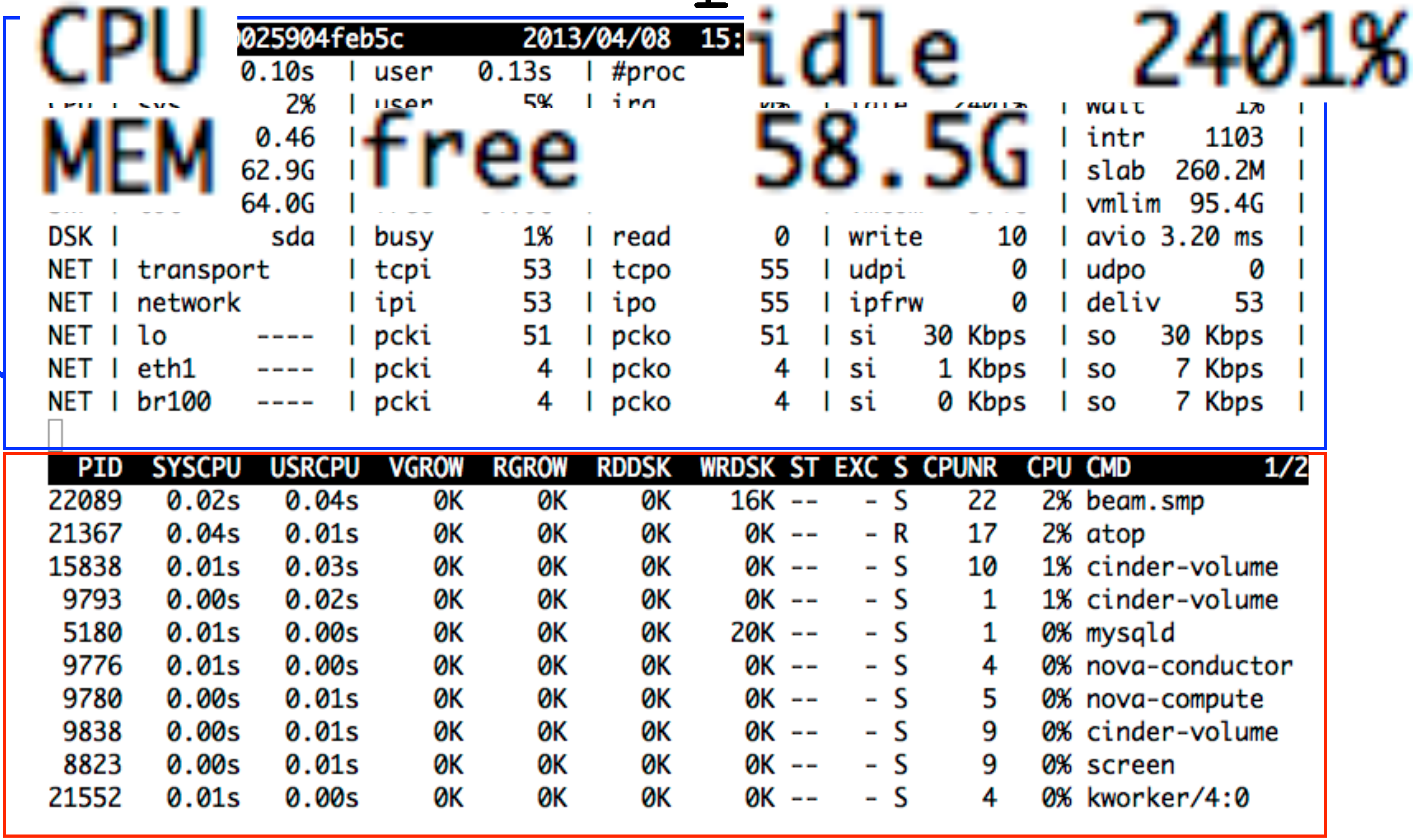
Per Process



```
025904feb5c          2013/04/08  15:
      0.10s  | user    0.13s  | #proc
         2%  | user       5%  | irq        0%  | idle   2401%  | wait      1%  |
      0.46   | avg5    0.16   | avg15   0.15   | csw     2058   | intr    1103  |
     62.9G   | free   58.5G   | cache   1.8G   | buff  177.1M   | slab  260.2M  |
     64.0G   | free   64.0G   |                | vmcom   3.4G   | vmlim  95.4G  |
DSK |          sda  | busy       1%  | read       0  | write     10  | avio 3.20 ms  |
NET | transport     | tcpi      53   | tcpo      55  | udpi       0  | udpo       0  |
NET | network       | ipi       53   | ipo       55  | ipfrw      0  | deliv     53  |
NET | lo       ---- | pcki      51   | pcko      51  | si   30 Kbps  | so   30 Kbps  |
NET | eth1     ---- | pcki       4   | pcko       4  | si    1 Kbps  | so    7 Kbps  |
NET | br100    ---- | pcki       4   | pcko       4  | si    0 Kbps  | so    7 Kbps  |

  PID  SYSCPU  USRCPU  VGROW  RGROW  RDDSK  WRDSK ST EXC S CPUNR  CPU CMD          1/2
22089   0.02s   0.04s     0K     0K     0K    16K --   - S    22   2% beam.smp
21367   0.04s   0.01s     0K     0K     0K     0K --   - R    17   2% atop
15838   0.01s   0.03s     0K     0K     0K     0K --   - S    10   1% cinder-volume
 9793   0.00s   0.02s     0K     0K     0K     0K --   - S     1   1% cinder-volume
 5180   0.01s   0.00s     0K     0K     0K    20K --   - S     1   0% mysqld
 9776   0.01s   0.00s     0K     0K     0K     0K --   - S     4   0% nova-conductor
 9780   0.00s   0.01s     0K     0K     0K     0K --   - S     5   0% nova-compute
 9838   0.00s   0.01s     0K     0K     0K     0K --   - S     9   0% cinder-volume
 8823   0.00s   0.01s     0K     0K     0K     0K --   - S     9   0% screen
21552   0.01s   0.00s     0K     0K     0K     0K --   - S     4   0% kworker/4:0
```

atop



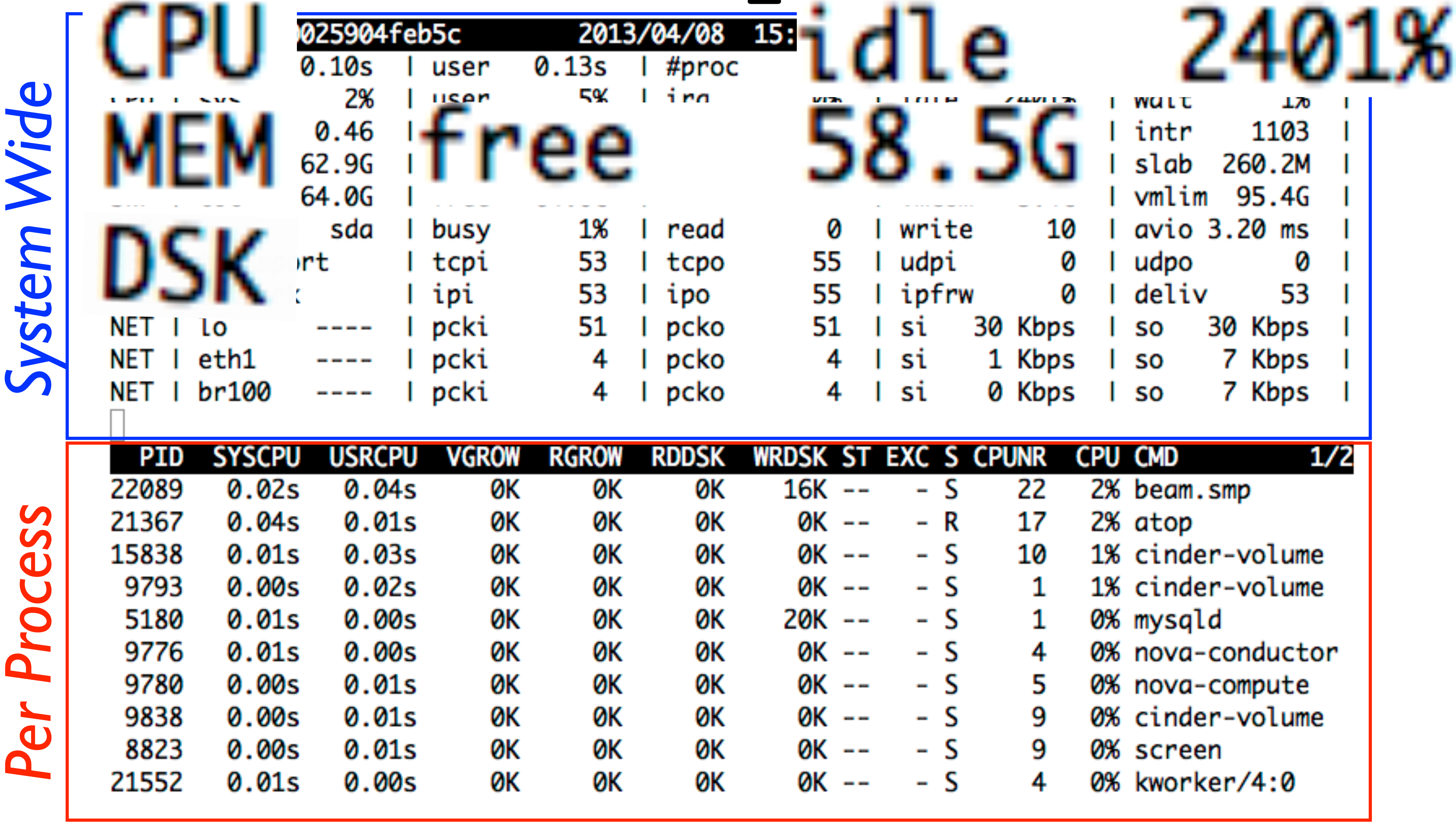
System Wide

```
CPU   025904feb5c        2013/04/08  15: idle         2401%
            0.10s  | user    0.13s  | #proc
               2%  | user       5%  | irq                             | wait        1%  |
MEM          0.46  | free                         58.5G               | intr      1103  |
            62.9G  |                                                  | slab    260.2M  |
            64.0G  |                                                  | vmlim    95.4G  |
DSK |         sda  | busy       1%  | read        0  | write      10  | avio 3.20 ms  |
NET | transport    | tcpi       53  | tcpo       55  | udpi        0  | udpo        0  |
NET | network      | ipi        53  | ipo        55  | ipfrw       0  | deliv      53  |
NET | lo      ---- | pcki       51  | pcko       51  | si   30 Kbps  | so   30 Kbps  |
NET | eth1    ---- | pcki        4  | pcko        4  | si    1 Kbps  | so    7 Kbps  |
NET | br100   ---- | pcki        4  | pcko        4  | si    0 Kbps  | so    7 Kbps  |
```

Per Process

```
  PID  SYSCPU  USRCPU  VGROW  RGROW  RDDSK  WRDSK ST EXC S CPUNR  CPU CMD          1/2
22089   0.02s   0.04s     0K     0K     0K    16K --   - S    22   2% beam.smp
21367   0.04s   0.01s     0K     0K     0K     0K --   - R    17   2% atop
15838   0.01s   0.03s     0K     0K     0K     0K --   - S    10   1% cinder-volume
 9793   0.00s   0.02s     0K     0K     0K     0K --   - S     1   1% cinder-volume
 5180   0.01s   0.00s     0K     0K     0K    20K --   - S     1   0% mysqld
 9776   0.01s   0.00s     0K     0K     0K     0K --   - S     4   0% nova-conductor
 9780   0.00s   0.01s     0K     0K     0K     0K --   - S     5   0% nova-compute
 9838   0.00s   0.01s     0K     0K     0K     0K --   - S     9   0% cinder-volume
 8823   0.00s   0.01s     0K     0K     0K     0K --   - S     9   0% screen
21552   0.01s   0.00s     0K     0K     0K     0K --   - S     4   0% kworker/4:0
```

atop



System Wide

```
CPU   025904feb5c           2013/04/08  15:  idle        2401%
          0.10s  | user     0.13s  | #proc
             2%  | user        5%  | irq                             | wait       1%  |
MEM        0.46  | free                      58.5G                   | intr     1103  |
          62.9G  |                                                   | slab   260.2M  |
          64.0G  |                                                   | vmlim   95.4G  |
DSK         sda  | busy        1%  | read        0  | write     10   | avio 3.20 ms  |
       rt        | tcpi        53  | tcpo       55  | udpi       0   | udpo       0  |
                 | ipi         53  | ipo        55  | ipfrw      0   | deliv     53  |
NET | lo    ---- | pcki        51  | pcko       51  | si   30 Kbps   | so   30 Kbps  |
NET | eth1  ---- | pcki         4  | pcko        4  | si    1 Kbps   | so    7 Kbps  |
NET | br100 ---- | pcki         4  | pcko        4  | si    0 Kbps   | so    7 Kbps  |
```

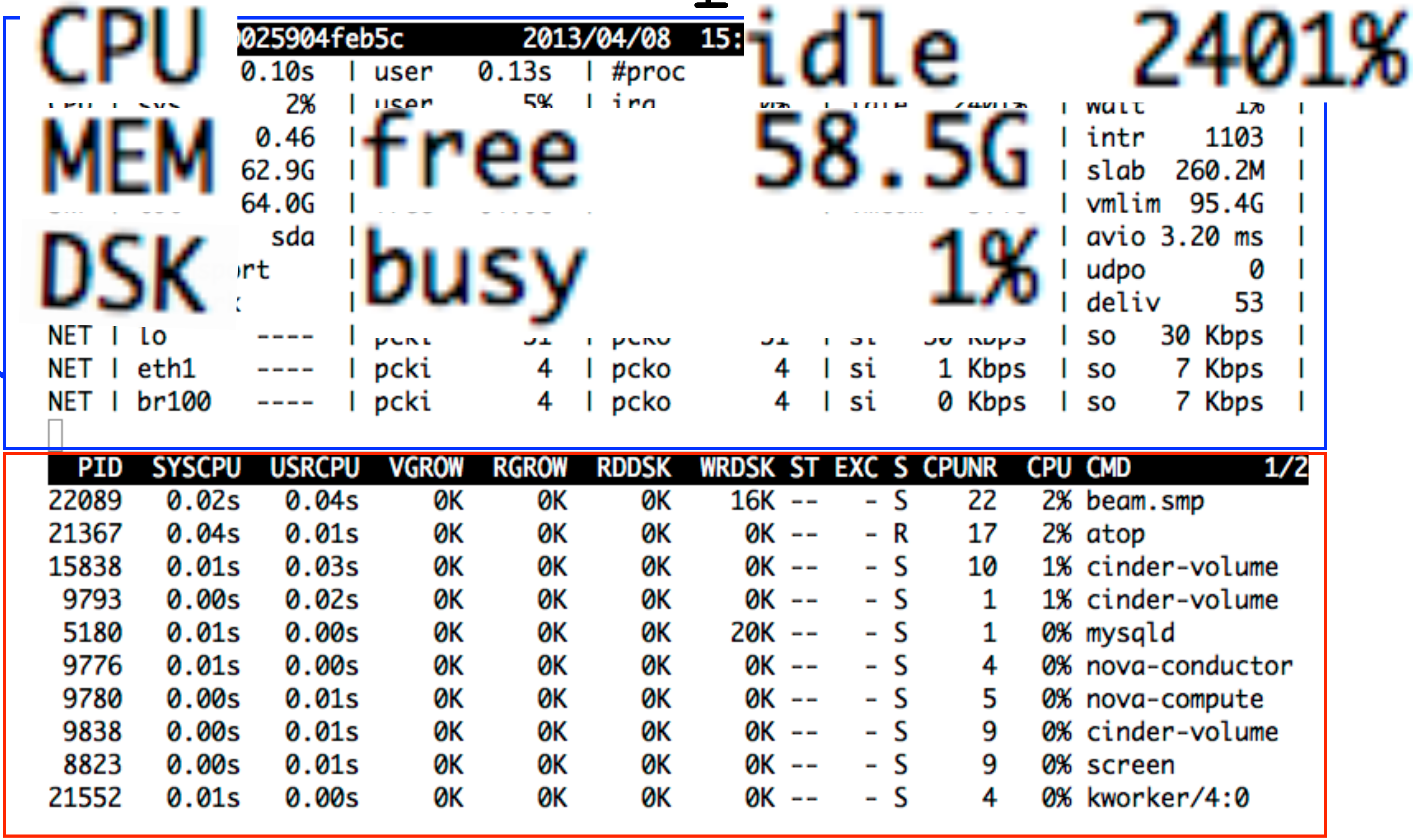
Per Process

```
  PID  SYSCPU  USRCPU  VGROW  RGROW  RDDSK  WRDSK  ST  EXC  S  CPUNR  CPU  CMD              1/2
22089   0.02s   0.04s     0K     0K     0K    16K  --     -  S     22   2%  beam.smp
21367   0.04s   0.01s     0K     0K     0K     0K  --     -  R     17   2%  atop
15838   0.01s   0.03s     0K     0K     0K     0K  --     -  S     10   1%  cinder-volume
 9793   0.00s   0.02s     0K     0K     0K     0K  --     -  S      1   1%  cinder-volume
 5180   0.01s   0.00s     0K     0K     0K    20K  --     -  S      1   0%  mysqld
 9776   0.01s   0.00s     0K     0K     0K     0K  --     -  S      4   0%  nova-conductor
 9780   0.00s   0.01s     0K     0K     0K     0K  --     -  S      5   0%  nova-compute
 9838   0.00s   0.01s     0K     0K     0K     0K  --     -  S      9   0%  cinder-volume
 8823   0.00s   0.01s     0K     0K     0K     0K  --     -  S      9   0%  screen
21552   0.01s   0.00s     0K     0K     0K     0K  --     -  S      4   0%  kworker/4:0
```

atop

System Wide

Per Process



```
CPU   025904feb5c          2013/04/08  15:    idle        2401%
       0.10s  | user    0.13s  | #proc
          2%  | user       5%  | irq                           | wait      1%  |
MEM    0.46  | free                          58.5G   | intr    1103  |
      62.9G  |                                       | slab  260.2M  |
      64.0G  |                                       | vmlim  95.4G  |
DSK     sda  | busy                            1%    | avio 3.20 ms  |
         rt  |                                       | udpo       0  |
             |                                       | deliv     53  |
NET | lo      ----  |                                  | so   30 Kbps  |
NET | eth1    ----  | pcki      4  | pcko      4  | si    1 Kbps  | so    7 Kbps  |
NET | br100   ----  | pcki      4  | pcko      4  | si    0 Kbps  | so    7 Kbps  |

  PID  SYSCPU  USRCPU  VGROW  RGROW  RDDSK  WRDSK ST EXC S CPUNR  CPU CMD          1/2
22089   0.02s   0.04s     0K     0K     0K    16K --   - S    22   2% beam.smp
21367   0.04s   0.01s     0K     0K     0K     0K --   - R    17   2% atop
15838   0.01s   0.03s     0K     0K     0K     0K --   - S    10   1% cinder-volume
 9793   0.00s   0.02s     0K     0K     0K     0K --   - S     1   1% cinder-volume
 5180   0.01s   0.00s     0K     0K     0K    20K --   - S     1   0% mysqld
 9776   0.01s   0.00s     0K     0K     0K     0K --   - S     4   0% nova-conductor
 9780   0.00s   0.01s     0K     0K     0K     0K --   - S     5   0% nova-compute
 9838   0.00s   0.01s     0K     0K     0K     0K --   - S     9   0% cinder-volume
 8823   0.00s   0.01s     0K     0K     0K     0K --   - S     9   0% screen
21552   0.01s   0.00s     0K     0K     0K     0K --   - S     4   0% kworker/4:0
```

Hardware Contention?

Hardware Contention?

- Sample every 2s using `atop -w log 2`

Hardware Contention?

- Sample every 2s using `atop -w log 2`
- HW utilization for *N*=20:

Hardware Contention?

- Sample every 2s using `atop -w log 2`
- HW utilization for *N*=20:

Resource	Metric
RAM	% Used
CPU	% Time Busy
Disk	% Time Busy

Hardware Contention?

- Sample every 2s using `atop -w log 2`
- HW utilization for *N*=20:

Resource	Metric	Median	Max
RAM	% Used	9	11
CPU	% Time Busy	14	55
Disk	% Time Busy	9	80

Hardware Contention?

- Sample every 2s using `atop -w log 2`
- HW utilization for *N*=20:

Resource	Metric	Median	Max
RAM	% Used	9	11
CPU	% Time Busy	14	55
Disk	% Time Busy	9	80

- Lots of capacity for parallelism

Hardware Contention?

- Sample every 2s using `atop -w log 2`
- HW utilization for *N*=20:

Resource	Metric	Median	Max
RAM	% Used	9	11
CPU	% Time Busy	14	55
Disk	% Time Busy	9	80

- Lots of capacity for parallelism
  - Time to look at SW

Software Bottlenecks

Software Bottlenecks

- Anything that inhibits parallelism
  - Some kind of lock contention

Software Bottlenecks

- Anything that inhibits parallelism
  - Some kind of lock contention
- Hopefully easy to fix :-)
  - Many locking strategies exist

Software Bottlenecks

- Anything that inhibits parallelism
  - Some kind of lock contention
- Hopefully easy to fix :-)
  - Many locking strategies exist
- Identified using tracing
  - Let’s take a look

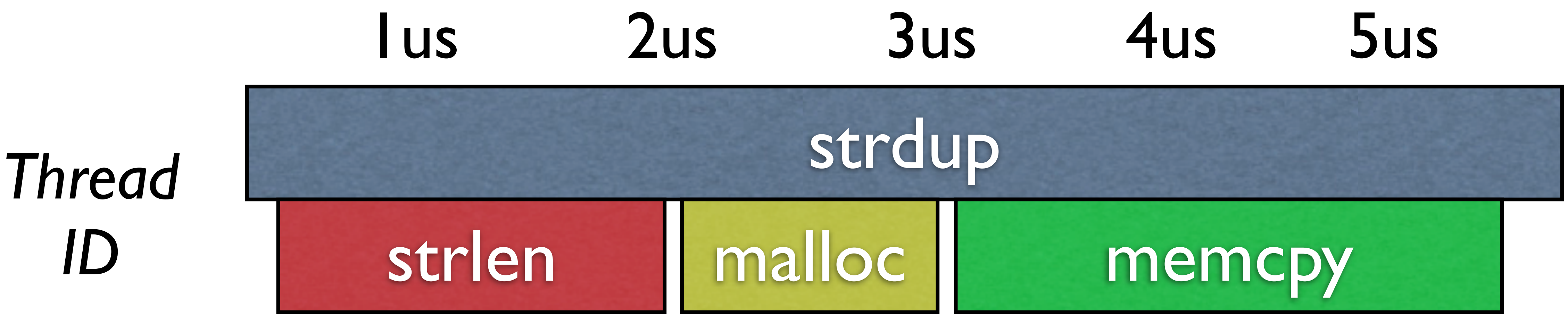
Tracing

Tracing

- Record events during application execution
  - e.g., Function entry & exit, lock acquisition

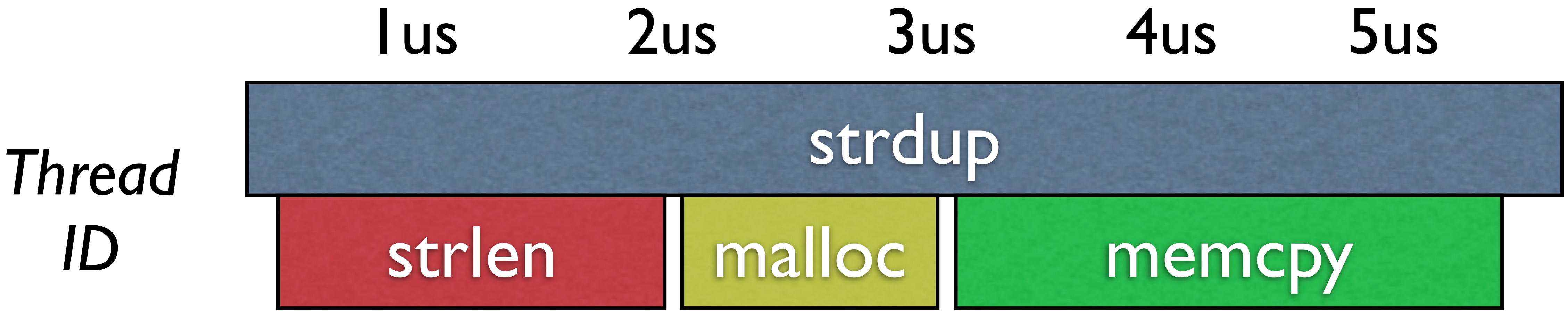
Tracing

- Record events during application execution
  - e.g., Function entry & exit, lock acquisition
- Visualized as stacked extents:



Tracing

- Record events during application execution
  - e.g., Function entry & exit, lock acquisition
- Visualized as stacked extents:



- Traces are usually pretty busy ...

Tracing OpenStack

Tracing OpenStack

- Added `@traced` to nova and quantum
  - Events on function call and return
  - Events before and after `lock()`
  - Outputs to trace-viewer format
    - Using Google Chrome? See about:tracing

Tracing OpenStack

- Added `@traced` to nova and quantum
  - Events on function call and return
  - Events before and after `lock()`
  - Outputs to trace-viewer format
    - Using Google Chrome? See about:tracing
- Repeat experiments with tracing on and hunt for bottlenecks
  - Look for stretched extents

Hunting: Resource Accounting

Hunting:
Resource Accounting

- Resource Accounting
  - Enforces max RAM, VCPUs, etc. allocated
  - Global lock per compute node

Bottleneck: Resource Lock

Bottleneck: Resource Lock

- Can add 15s of serialization to VM creation

Bottleneck: Resource Lock

- Can add 15s of serialization to VM creation
- Slow because of RPC to conductor

Bottleneck: Resource Lock

- Can add 15s of serialization to VM creation
- Slow because of RPC to conductor
- Solution Part 1: Remove NOP updates
    - Reduces median creation time 10% when *N*=20

Bottleneck: Resource Lock

- Can add 15s of serialization to VM creation
- Slow because of RPC to conductor
- Solution Part 1: Remove NOP updates
  - Reduces median creation time 10% when *N*=20
- Solution Part 2: Coalesce RPCs
  - Future work

Hunting: Libvirt

Hunting: Libvirt

- libvirt starts qemu process, apparmor, etc.

Hunting: Libvirt

- libvirt starts qemu process, apparmor, etc.
- Global lock... can't fix this in OpenStack

Hunting: Libvirt

- libvirt starts qemu process, apparmor, etc.
- Global lock... can’t fix this in OpenStack
- Can we mitigate the problem?

Bottleneck: Libvirt

Bottleneck: Libvirt

- Many short calls (e.g., get hostname)
  - Become long calls due to global lock

Bottleneck: Libvirt

- Many short calls (e.g., get hostname)
  - Become long calls due to global lock
- Solution: avoid unnecessary calls
  - Down from 248 to 7
  - Reduces max creation time 20% when *N*=20

Hunting: Eventlet

Hunting: Eventlet

- eventlet's "green" threads are coroutines multiplexed on single native thread
  - You can't block in a green thread
  - Python's stdlib patched to yield instead of block
  - C libraries aren't patched

Hunting: Eventlet

- eventlet's "green" threads are coroutines multiplexed on single native thread
  - You can't block in a green thread
  - Python's stdlib patched to yield instead of block
  - C libraries aren't patched
- Pool of native threads to use blocking libs

Hunting: Eventlet

- eventlet's "green" threads are coroutines multiplexed on single native thread
  - You can't block in a green thread
  - Python's stdlib patched to yield instead of block
  - C libraries aren't patched
- Pool of native threads to use blocking libs
- Maybe there's more room for improvement

Bottleneck:
Eventlet Work Queues

Bottleneck: Eventlet Work Queues

- One work queue per worker thread

Bottleneck:
Eventlet Work Queues

- One work queue per worker thread
- Green-thread to work-queue map is fixed:

```
worker_idx = hash(gettid()) % \
             worker_count
work_queues[worker_idx].append(work)
```

Bottleneck: Eventlet Work Queues

- One work queue per worker thread
- Green-thread to work-queue map is fixed:

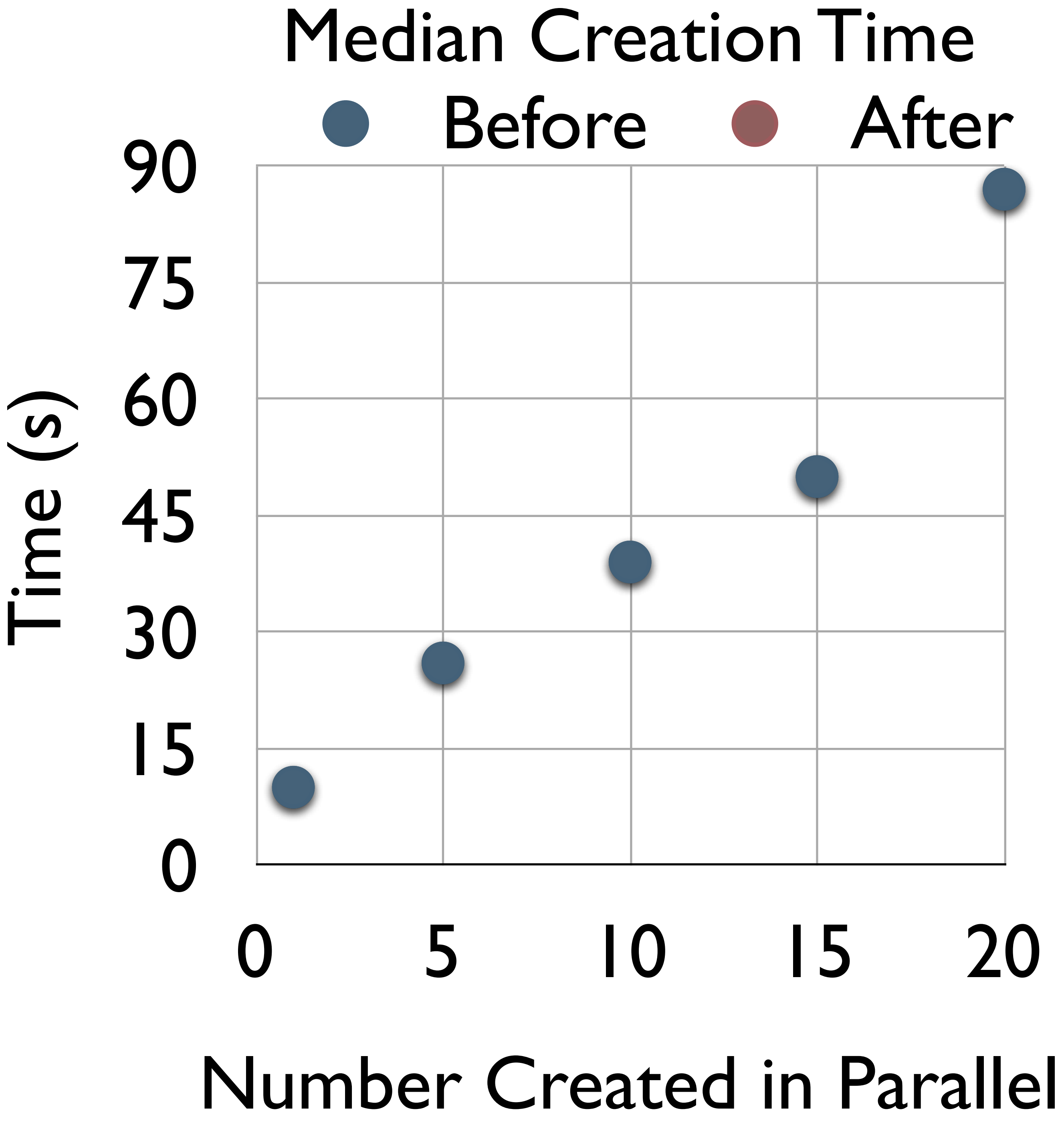
```
worker_idx = hash(gettid()) % \
             worker_count
work_queues[worker_idx].append(work)
```

- Solution: use a global work queue
  - Get to wait on libvirt lock sooner :'-(

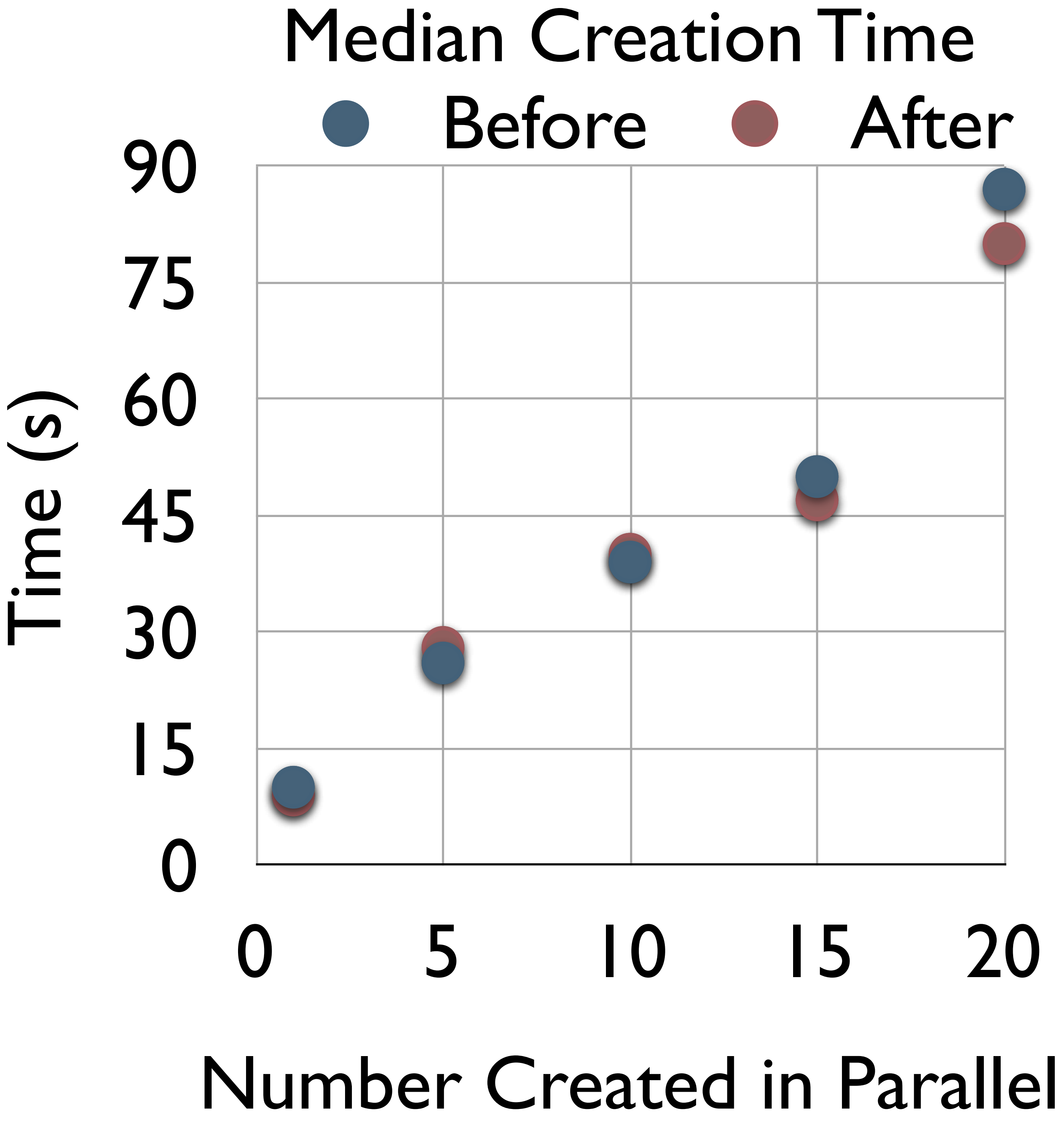
Results

● Before ● After

Results

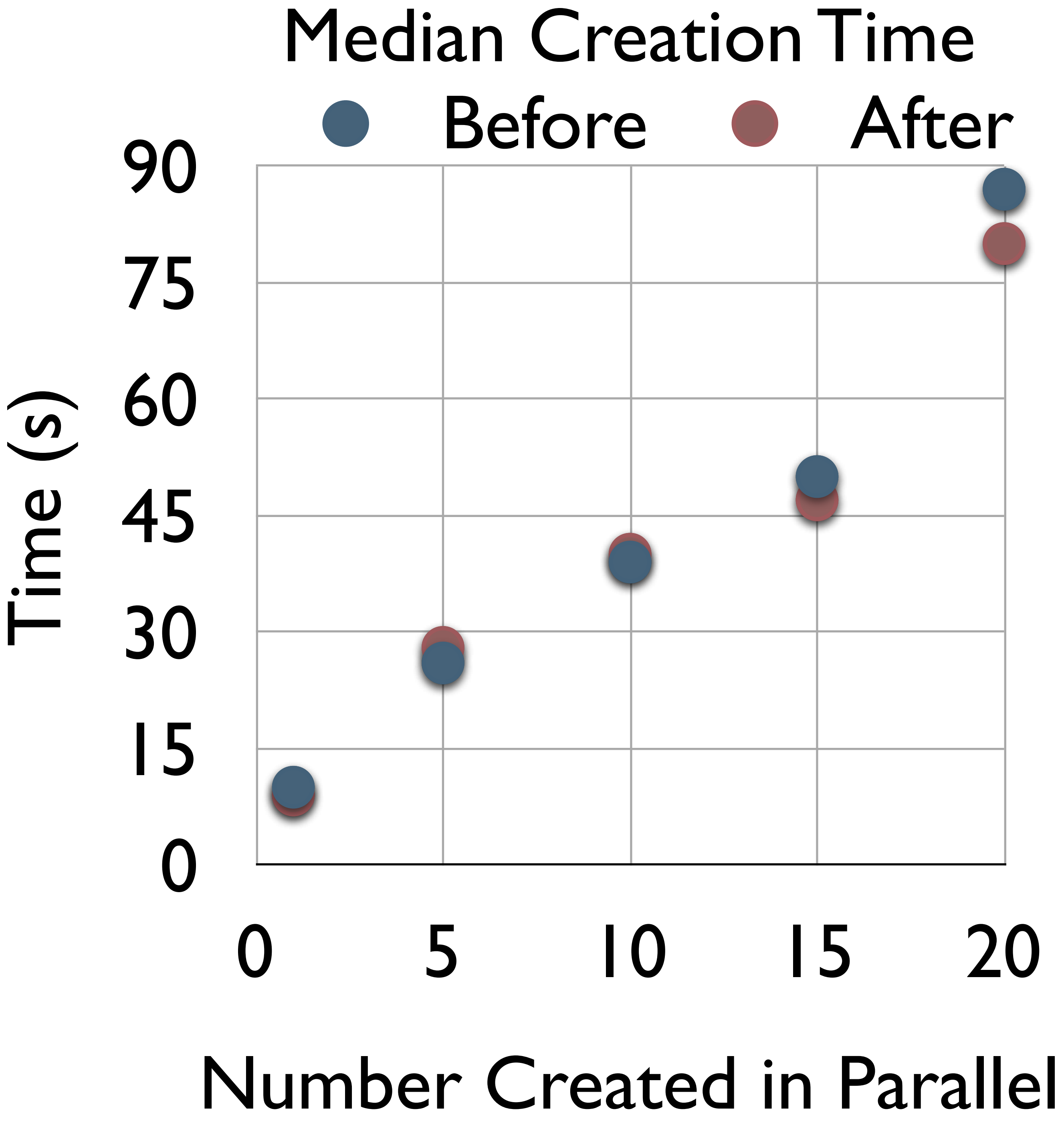

Median Creation Time
● Before
● After
90
75
60
45
30
15
0
Time (s)
0
5
10
15
20
Number Created in Parallel

Results


Median Creation Time
● Before
● After
90
75
60
45
30
15
0
Time (s)
0
5
10
15
20
Number Created in Parallel

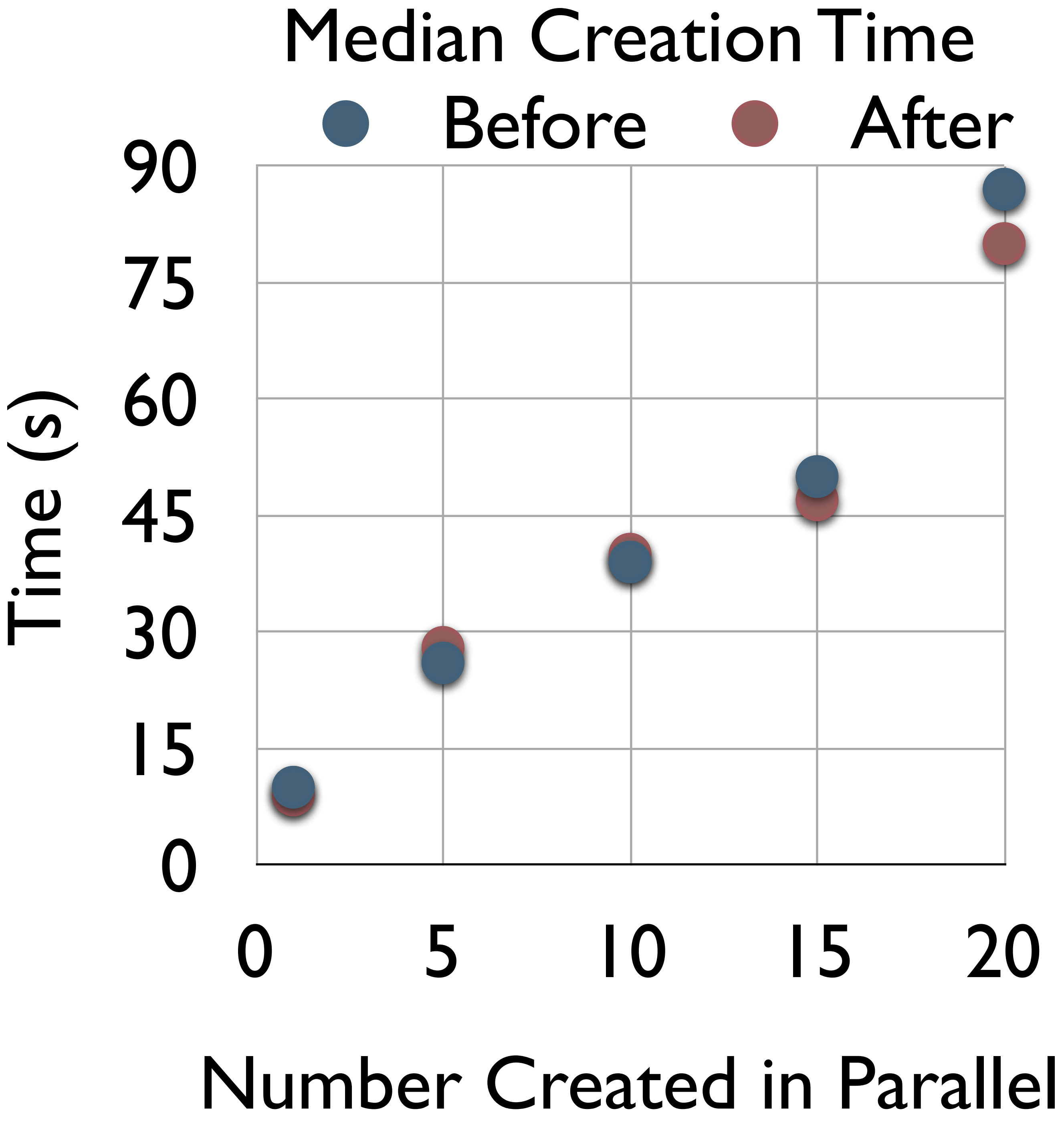
Results

- VM creation time:
  - Max 20% lower
  - Median 10% lower



Results

- VM creation time:
  - Max 20% lower
  - Median 10% lower
- Wait for libvirt sooner
  - On the bright side, once libvirt fixed, OpenStack has fewer bottlenecks



Conclusion

- Low VM creation time is good
  - Necessary for scaling
- VM Creation time scales poorly due to software contention
  - Bottlenecks in OpenStack code easily fixed
  - libvirt still a big bottleneck
- Tracing helps identify contention

Future Work

- Coalesce RPC updates to conductor
- Eliminate big qemu lock in libvirt
- Instrument other OpenStack services (glance, swift, cinder, etc.)
- Perform more experiments

Questions?

gridcentric

Peter Feiner

peter@gridcentric.com

github.com/peterfeiner/{nova,quantum}/tree/tracing