

從企業看 SA/NA 及經驗分享

小明(曾祺元)

2019.11.28

姓名：曾祺元

綽號：小明

現任：PIXNET SRE 組組長

經歷：

- 璞園建築團隊 IT 主管
- 遊戲基地(gamebase)資深系統工程師
- 資策會創研所資深工程師
- 台大生機所 90 級
- 交大機械系 89 級
 - 曾擔任 CCCA 及交大機械系網管



台灣最大社群網站

PIXNET 創立於2003年，2006年成立「優像數位媒體科技股份有限公司」，並於2007年加入城邦媒體控股集團。我們是一間以社群為核心的科技公司，旗下主要服務包含：痞客邦、PIXgoods、PIXmarketing、PIXinsight，透過創新的數據應用、多樣化社群服務，實現「Guide to SMART Life」企業核心價值。2018年，PIXNET 推出「全新痞客邦」加速興趣同好彼此凝聚及交流，並持續與產業各界結盟，踏實建構「社群共榮圈」願景。



PIXNET
Guide to SMART Life





國立交通大學

校園網路策進會

Campus Computer

Communication Association



Since 1992 May

打地基

蓋高樓

房屋買賣

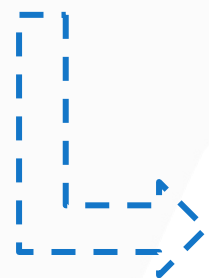
網路架構
系統架構

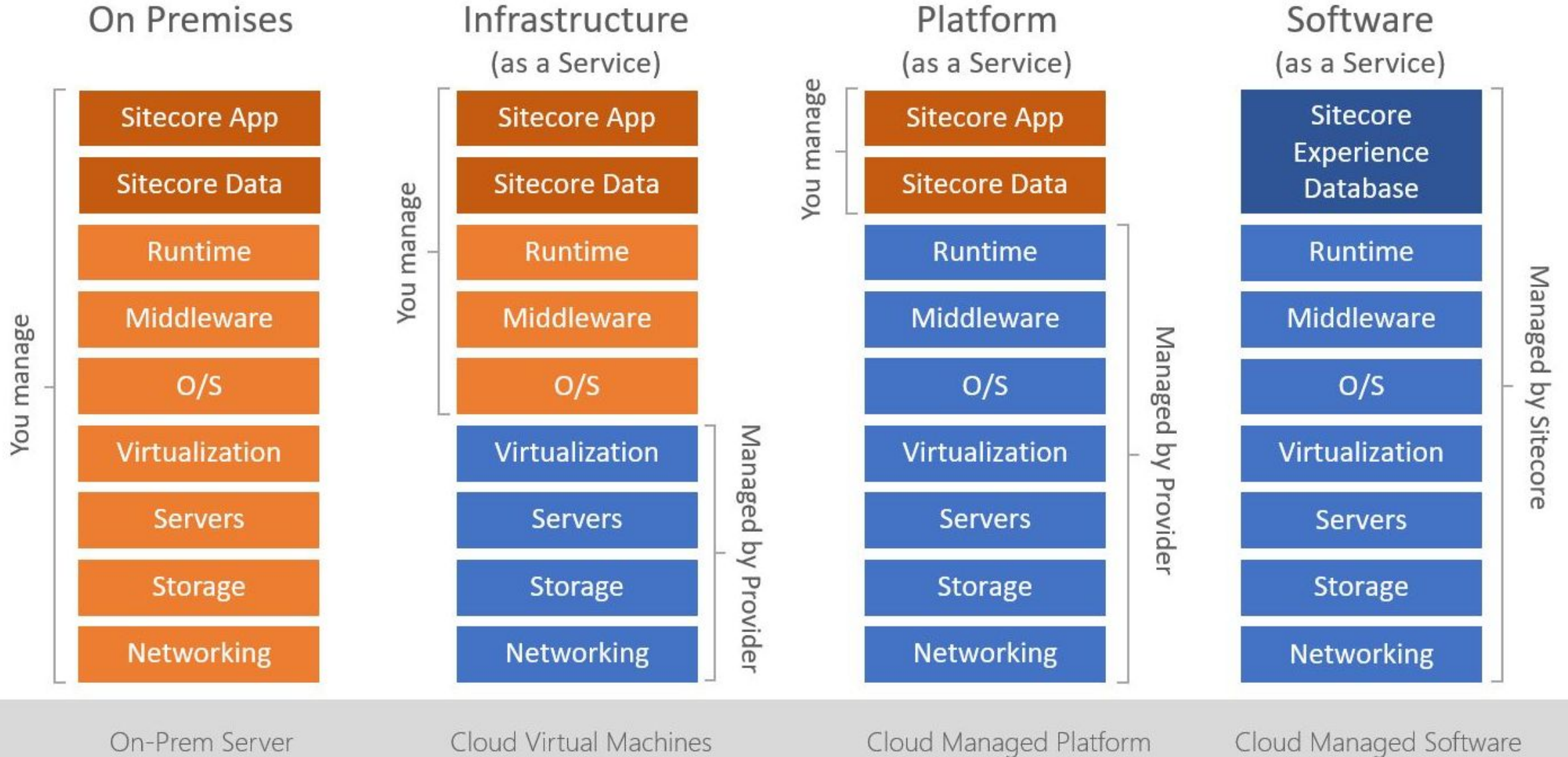
軟體開發

營運、維運

地端機房

雲端服務

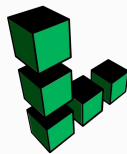


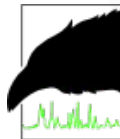
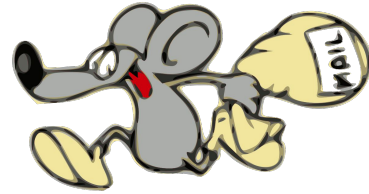


地端機房



雲端服務





SA / NA 基礎建設的根本

來談談一些維運的應用

值班

- 維運
- 穩定
- 救火

支援

- 協助 RD
- 建置環境

開發

- 自動化
- 新技術

ZFS + Percona (MySQL)

說說一些 ZFS 的補充

- ZFS is a combined **file system** and **logical volume manager** designed by Sun Microsystems.
- The ZFS file system is a file system that fundamentally **changes** the way file systems are administered, with **features** and **benefits** not found in other file systems available today. ZFS is robust, scalable, and easy to administer.

- Software Raid - recommend **HBA card**
- 128 bit filesystem
- no fsck - scrub / resilvering
- RAID-Z / mirror
- Snapshots

Platform

- Solaris / OpenSolaris
- macOS / FreeBSD
- FreeNAS / NAS4free / pfsense

<https://en.wikipedia.org/wiki/ZFS>

<https://zfsonlinux.org/>

http://wiki.lustre.org/ZFS_OSD_Hardware_Considerations

https://docs.oracle.com/cd/E26505_01/html/E37384/zfsover-2.html#scrolltoc

忽略硬碟排列順序

FreeBSD: ZFS + GEOM

Debian: ZFS + disk path / label

```
pool: storage
state: ONLINE
scan: scrub repaired 0B in 1h13m with 0 errors on Sun Nov 10 01:37:19 20
config:

NAME
storage
raidz1-0
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy2-lun-0-part1
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy3-lun-0-part1
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy4-lun-0-part1
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy5-lun-0-part1
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy6-lun-0-part1
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy7-lun-0-part1
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy8-lun-0-part1
logs
mirror-1
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy38-lun-0-part3
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy39-lun-0-part3
cache
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy0-lun-0-part1
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy1-lun-0-part1
spares
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy9-lun-0-part1
```

```
pool: storage
state: ONLINE
scan: scrub repaired 0 in 0 days 03:35:21 with 0 error
config:

NAME                STATE      READ WRITE CKSUM
storage             ONLINE    0     0     0
  raidz1-0          ONLINE    0     0     0
    gpt/storage0    ONLINE    0     0     0
    gpt/storage1    ONLINE    0     0     0
    gpt/storage2    ONLINE    0     0     0
    gpt/storage3    ONLINE    0     0     0
    gpt/storage4    ONLINE    0     0     0
    gpt/storage5    ONLINE    0     0     0
    gpt/storage6    ONLINE    0     0     0
  logs
    mirror-1        ONLINE    0     0     0
      gpt/write0     ONLINE    0     0     0
      gpt/write1     ONLINE    0     0     0
  cache
    gpt/read0        ONLINE    0     0     0
    gpt/read1        ONLINE    0     0     0
  spares
    gpt/spare0       AVAIL
```

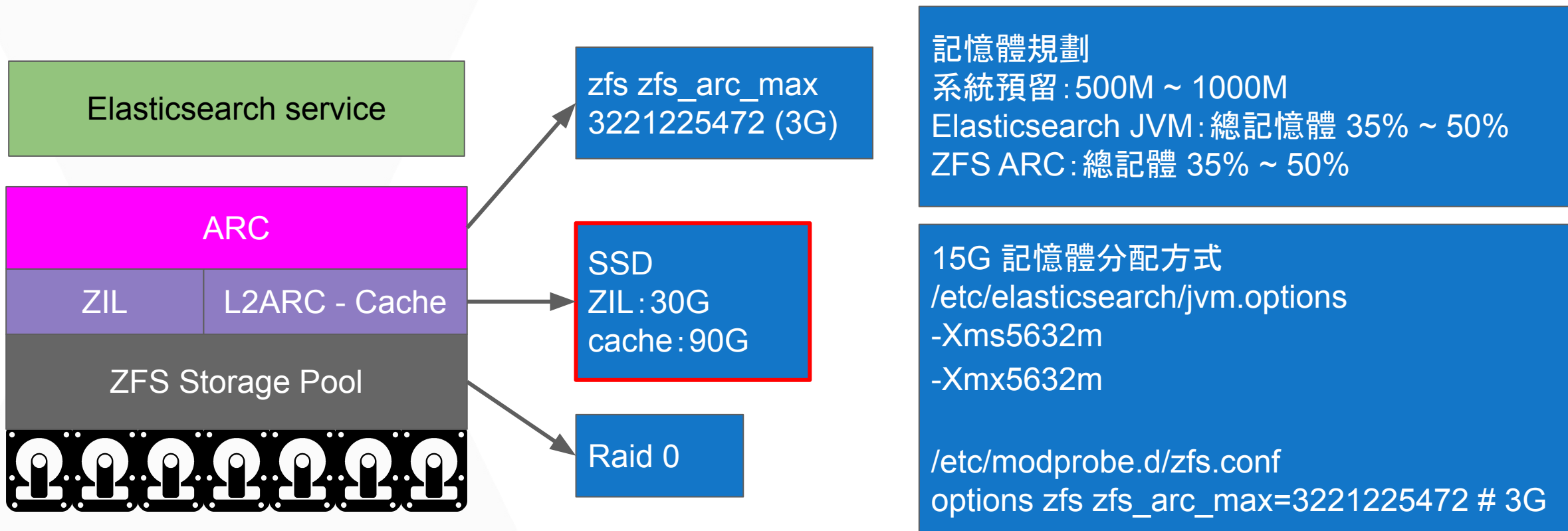
Root on ZFS 很好用
但需要和儲存資料分離

提升 IOPS
移用 SSD 當 ZIL & Cache

故事

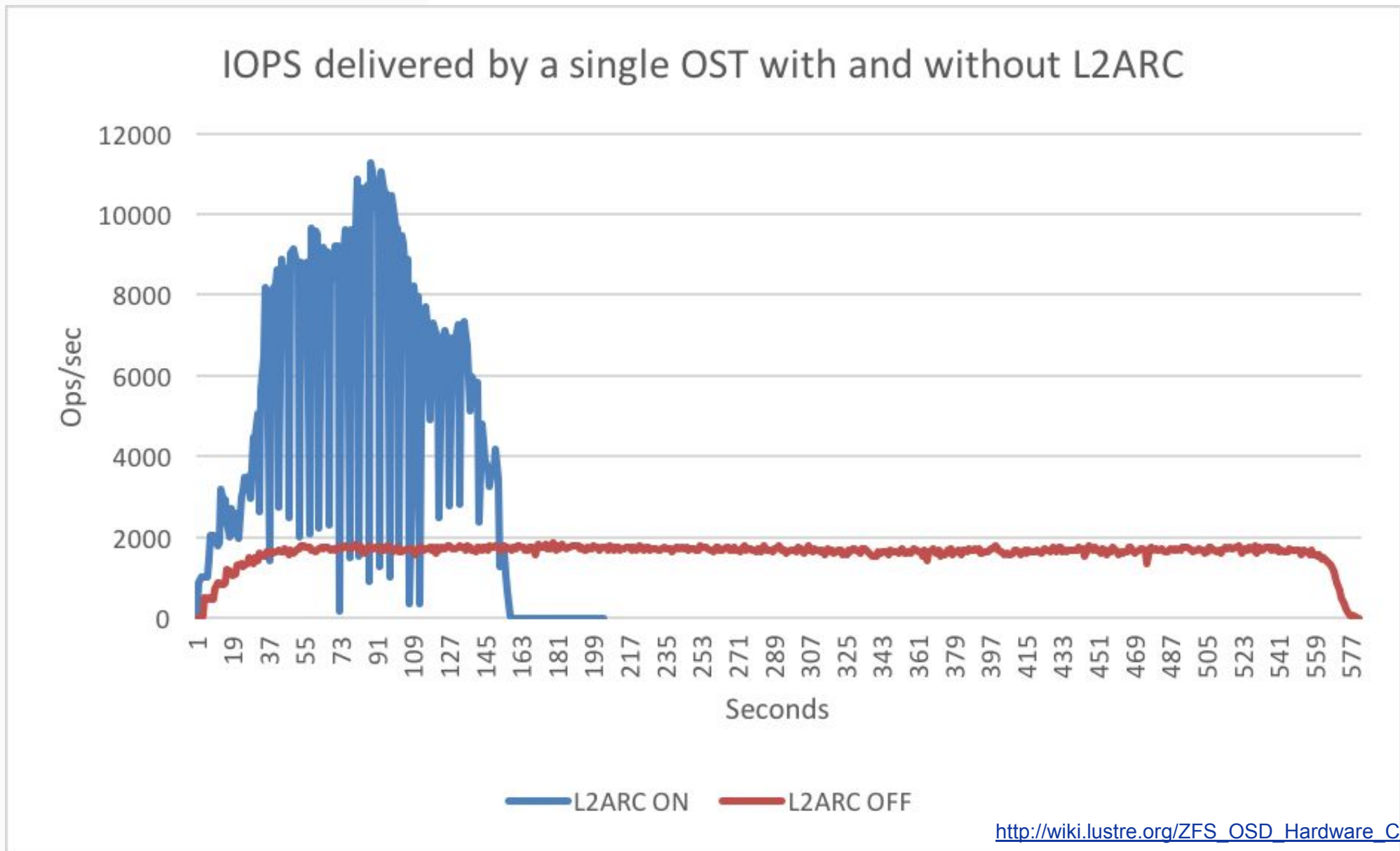
- 曾在過年時發生 Storage (ZFS) 因故無法登入使用
- 年初三回公司處理，重開機再也找不到 root partition
- 各種方法都救不回來 root partition
- 最後用額外的硬碟當 OS 後 import ZFS 救回

降低機器負擔 (Percona 為例)
讀寫分離、讀寫比例分配

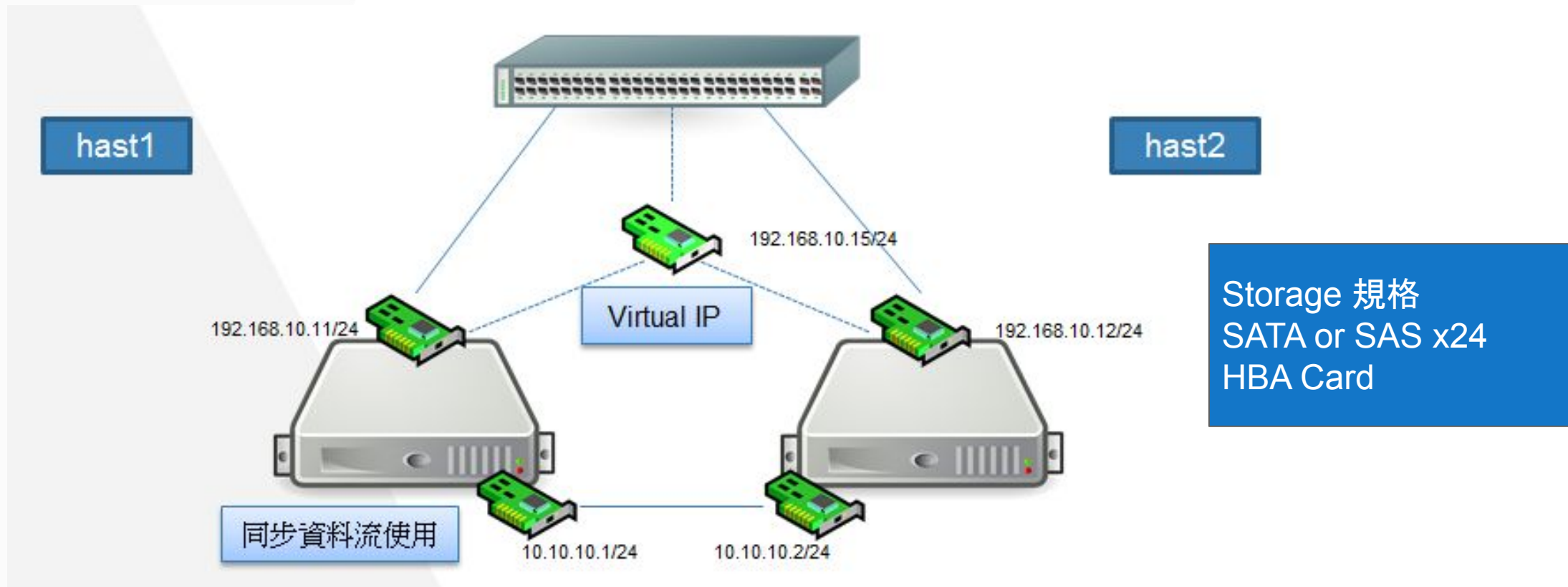


L2ARC

- 若有獨立 (SSD) Cache 則稱為 SLOG (Separate ZFS Intent Log, SLOG)
- 若沒有獨立 Cache 則由所有 (virtual devices, vdevs) 分擔 ZIL 功能

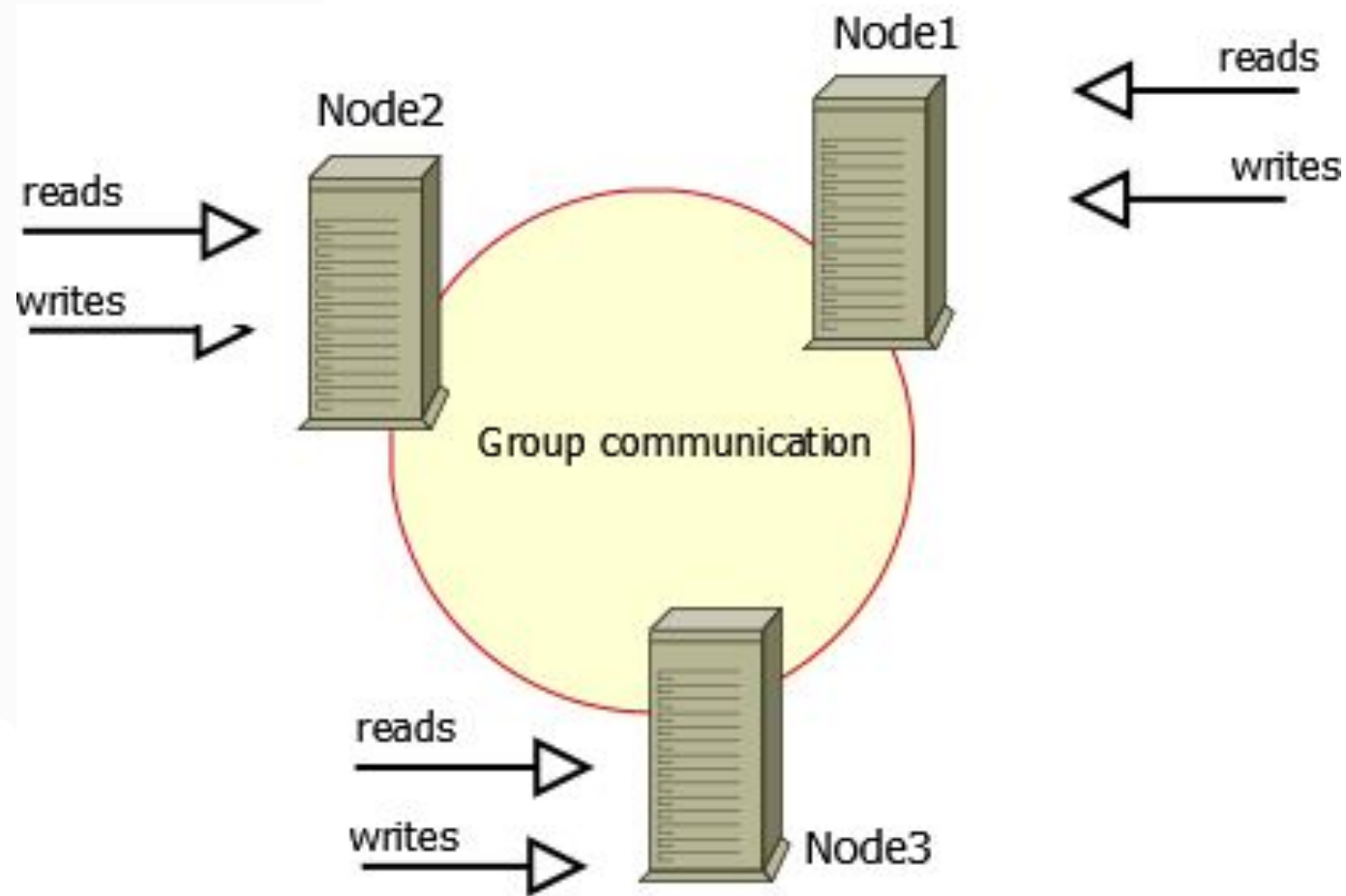


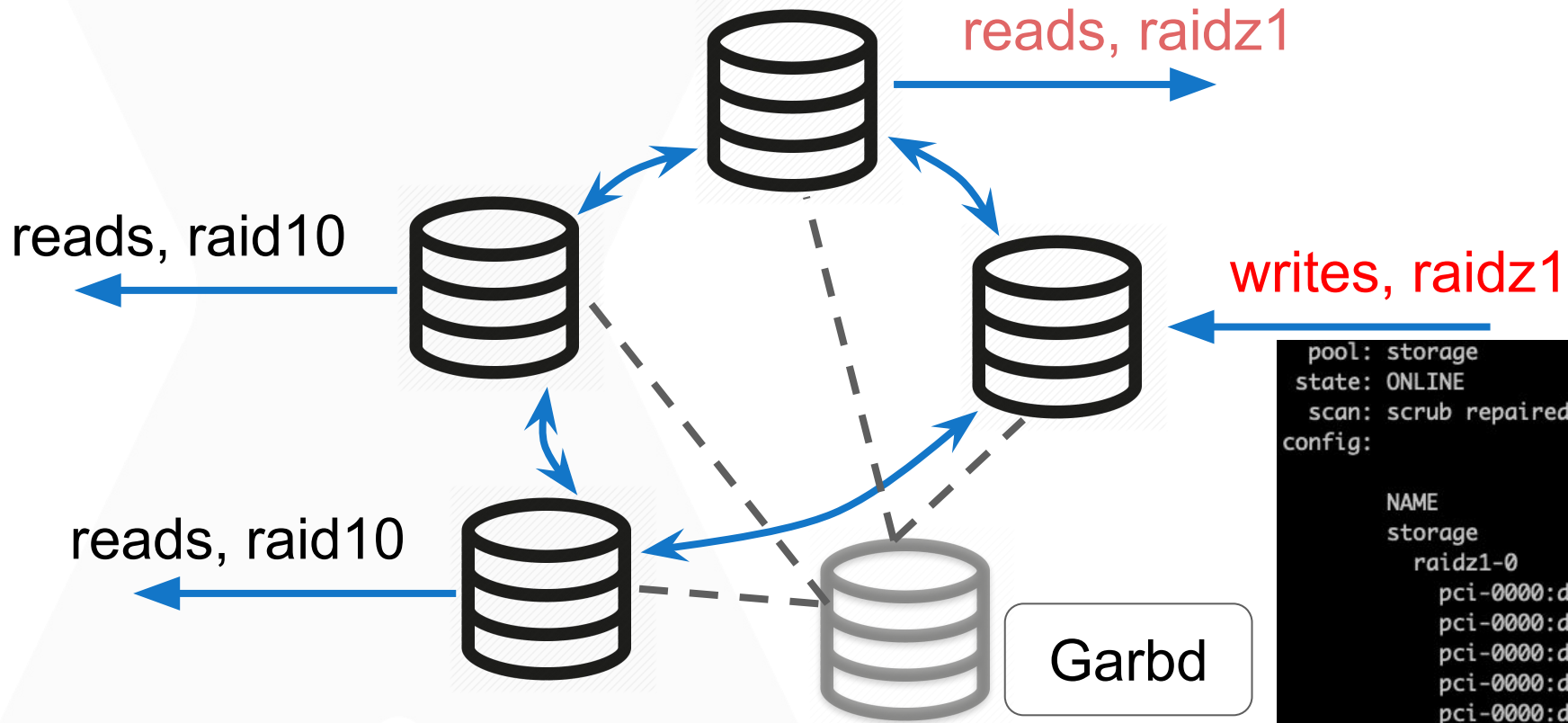
FreeBSD + HAST + CARP + ZFS = 超好用



- VMWare GuestOS 在 ZFS Failover 時可不中斷服務
- ZFS Failover 時間約 78s

回來談談 Percona (MySQL)





```

pool: storage
state: ONLINE
scan: scrub repaired 0B in 1h13m with 0 errors on Sun Nov 10 01:37:19 20
config:

NAME
storage
raidz1-0
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy2-lun-0-part1
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy3-lun-0-part1
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy4-lun-0-part1
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy5-lun-0-part1
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy6-lun-0-part1
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy7-lun-0-part1
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy8-lun-0-part1

logs
mirror-1
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy38-lun-0-part3
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy39-lun-0-part3

cache
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy0-lun-0-part1
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy1-lun-0-part1

spares
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy9-lun-0-part1
    
```

```
top - 19:51:04 up 397 days, 4:03, 2 users, load average: 5.71, 5.45, 5.06
Tasks: 341 total, 1 running, 340 sleeping, 0 stopped, 0 zombie
%Cpu(s): 1.9 us, 0.4 sy, 0.0 ni, 85.7 id, 11.9 wa, 0.0 hi, 0.1 si, 0.0 st
KiB Mem: 65999196 total, 65541344 used, 457852 free, 93400 buffers
KiB Swap: 62498812 total, 1237936 used, 61260876 free. 6360104 cached Mem
```

PID	USER	PR	NI	VIRT	RES	SHR	S	%CPU	%MEM	TIME+	COMMAND
2246	mysql	20	0	85.895g	0.059t	4.868g	S	73.0	95.5	137547:19	mysqld

```
avg-cpu:  %user   %nice %system %iowait  %steal   %idle
           2.14    0.00    0.41    7.49    0.00   89.97
```

```
Device:            rrqm/s   wrqm/s     r/s     w/s    rkB/s     kB/s avgrq-sz avgqu-sz   await r_await w_await  svctm  %util
sda                 0.00     0.00  539.00  34.00  3704.00   152.50   13.46     3.17    5.41   5.74   0.12   1.75  100.00
```

```
top - 21:45:20 up 84 days, 13:26, 2 users, load average: 6.85, 7.12, 7.00
Tasks: 471 total, 1 running, 470 sleeping, 0 stopped, 0 zombie
%Cpu(s): 13.4 us, 3.0 sy, 0.0 ni, 80.8 id, 0.0 wa, 0.0 hi, 2.9 si, 0.0 st
MiB Mem : 128596.4 total, 13225.1 free, 113380.7 used, 1990.6 buff/cache
MiB Swap: 61440.0 total, 61235.7 free, 204.2 used. 14113.3 avail Mem
```

PID	USER	PR	NI	VIRT	RES	SHR	S	%CPU	%MEM	TIME+	COMMAND
18092	mysql	20	0	62.3g	49.4g	154780	S	560.6	39.4	376792:21	mysqld

```
avg-cpu:  %user   %nice %system %iowait  %steal   %idle
           17.39    0.00    5.95    0.83    0.00    75.83
```

Device	tps	kB/s	rqm/s	await	aqu-sz	areq-sz	%util
sdc	115.00	1942.00	0.00	5.31	0.62	16.89	31.60
sdf	113.00	1942.50	0.00	5.35	0.61	17.19	32.40
sde	116.00	1942.00	0.00	5.25	0.61	16.74	31.60
sdh	114.00	1941.50	0.00	5.81	0.67	17.03	34.80
sdj	0.00	0.00	0.00	0.00	0.00	0.00	0.00
sdi	120.00	1941.00	0.00	4.94	0.60	16.18	30.40
sda	0.00	0.00	0.00	0.00	0.00	0.00	0.00
sdb	0.00	0.00	0.00	0.00	0.00	0.00	0.00
sdk	0.00	0.00	0.00	0.00	0.00	0.00	0.00
sdl	0.00	0.00	0.00	0.00	0.00	0.00	0.00
sdd	114.00	1939.50	0.00	4.99	0.60	17.01	30.80
sdg	117.00	1944.50	0.00	5.16	0.61	16.62	32.00
md0	0.00	0.00	0.00	0.00	0.00	0.00	0.00
zd0	0.00	0.00	0.00	0.00	0.00	0.00	0.00

```
pool: storage
state: ONLINE
scan: scrub repaired 0B in 1h13m with 0 errors on Sun Nov 10 01:37:19 2019
config:
```

NAME	STATE	READ	WRITE	CKSUM
storage	ONLINE	0	0	0
raidz1-0	ONLINE	0	0	0
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy2-lun-0-part1	ONLINE	0	0	0
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy3-lun-0-part1	ONLINE	0	0	0
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy4-lun-0-part1	ONLINE	0	0	0
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy5-lun-0-part1	ONLINE	0	0	0
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy6-lun-0-part1	ONLINE	0	0	0
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy7-lun-0-part1	ONLINE	0	0	0
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy8-lun-0-part1	ONLINE	0	0	0
logs				
mirror-1	ONLINE	0	0	0
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy38-lun-0-part3	ONLINE	0	0	0
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy39-lun-0-part3	ONLINE	0	0	0
cache				
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy0-lun-0-part1	ONLINE	0	0	0
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy1-lun-0-part1	ONLINE	0	0	0
spares				
pci-0000:d8:00.0-sas-exp0x500304801f1546ff-phy9-lun-0-part1	AVAIL			

```
errors: No known data errors
```



```

top - 21:53:33 up 84 days, 13:26, 1 user, load average: 4.81, 5.24, 5.41
Tasks: 457 total, 1 running, 456 sleeping, 0 stopped, 0 zombie
%Cpu(s): 11.4 us, 2.7 sy, 0.0 ni, 83.8 id, 0.1 wa, 0.0 hi, 1.9 si, 0.0 st
MiB Mem : 64308.3 total, 14979.9 free, 47656.3 used, 1672.2 buff/cache
MiB Swap: 61440.0 total, 61001.5 free, 438.5 used. 15930.8 avail Mem

```

PID	USER	PR	NI	VIRT	RES	SHR	S	%CPU	%MEM	TIME+	COMMAND
4632	mysql	20	0	38.6g	25.5g	145284	S	479.1	40.6	321266:35	mysqld

```

avg-cpu:  %user   %nice %system %iowait  %steal   %idle
           9.65    0.00   4.59    1.10    0.00    84.65

Device            tps    kB/s    rqm/s    await    aqu-sz    areq-sz    %util
sdb                85.00   4493.50   0.00     4.93     0.44     52.86    22.00
sda                86.00   4493.50   0.00     5.16     0.44     52.25    22.40
sdc                86.00   4493.50   1.00     5.45     0.46     52.25    23.20
sdd                85.00   4493.50   0.00     5.56     0.47     52.86    24.00
sdf               111.00   6141.00   0.00     6.29     0.69     55.32    34.80
sdg               111.00   6141.00   0.00     5.95     0.66     55.32    32.80
sde               110.00   6141.00   0.00     6.36     0.71     55.83    36.00
sdh               112.00   6141.00   0.00     6.02     0.67     54.83    34.80
sdi                7.00    189.00   2.00     0.86     0.00     27.00     0.00
sdj                6.00    145.00   2.00     1.17     0.00     24.17     0.00
md0                4.00    16.00   0.00     0.00     0.00     4.00     0.00
zd0                0.00     0.00   0.00     0.00     0.00     0.00     0.00

```

```

pool: storage
state: ONLINE
scan: scrub repaired 0B in 0h38m with 0 errors on Sun Nov 10 01:02:44 2019
config:

NAME                                STATE      READ WRITE CKSUM
storage                              ONLINE    0    0    0
mirror-0
  pci-0000:82:00.0-scsi-0:0:69:0-part1 ONLINE    0    0    0
  pci-0000:82:00.0-scsi-0:0:70:0-part1 ONLINE    0    0    0
  pci-0000:82:00.0-scsi-0:0:71:0-part1 ONLINE    0    0    0
  pci-0000:82:00.0-scsi-0:0:72:0-part1 ONLINE    0    0    0
mirror-1
  pci-0000:82:00.0-scsi-0:0:73:0-part1 ONLINE    0    0    0
  pci-0000:82:00.0-scsi-0:0:74:0-part1 ONLINE    0    0    0
  pci-0000:82:00.0-scsi-0:0:75:0-part1 ONLINE    0    0    0
  pci-0000:82:00.0-scsi-0:0:76:0-part1 ONLINE    0    0    0
logs
mirror-2
  pci-0000:00:1f.2-ata-1-part3      ONLINE    0    0    0
  pci-0000:00:1f.2-ata-2-part3      ONLINE    0    0    0
cache
  pci-0000:00:1f.2-ata-1-part4      ONLINE    0    0    0
  pci-0000:00:1f.2-ata-2-part4      ONLINE    0    0    0

errors: No known data errors

```

環境

CPU: 8 Cores 以上

RAM: 64G 以上

HDD: SAS 300G x8 (600G) 以上

SSD: 120G x 2

ZFS 參數

atime=off

checksum=fletcher4

setuid=off

exec=off

devices=off

sync=disabled for mysql datadir and binglog dir

ZIL: 30G (mirror)

cache: 90G stripe

ARC: 20G

Percona 參數

innodb_buffer_pool_size = 20G

innodb_thread_concurrency = 32: 以 CPU 數量而定, cat /proc/cpuinfo | grep proc | wc -l

innodb_read_io_threads = 28

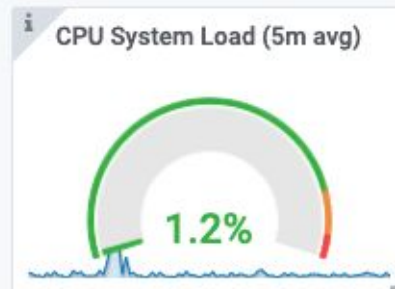
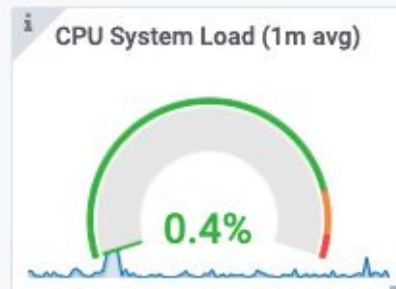
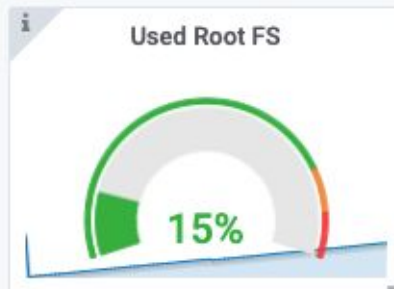
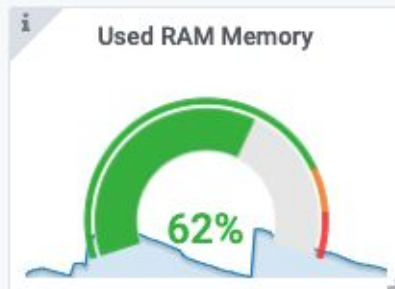
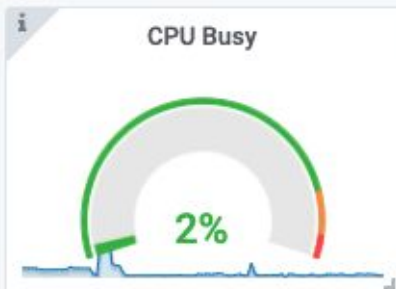
innodb_write_io_threads = 4

: io theads 則依讀與寫的量比例分配

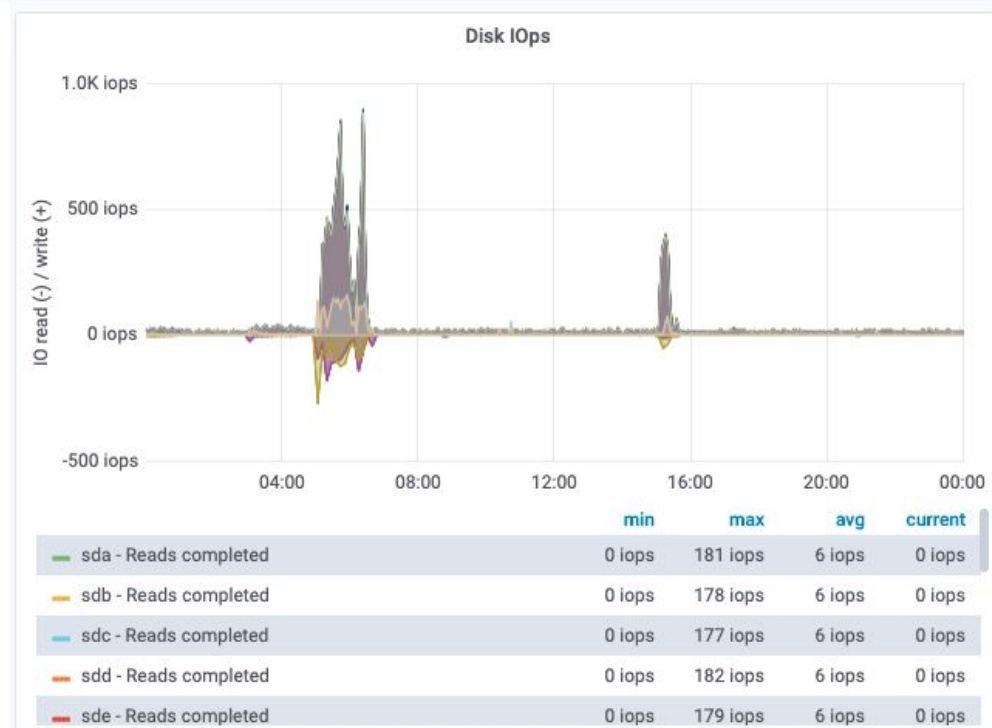
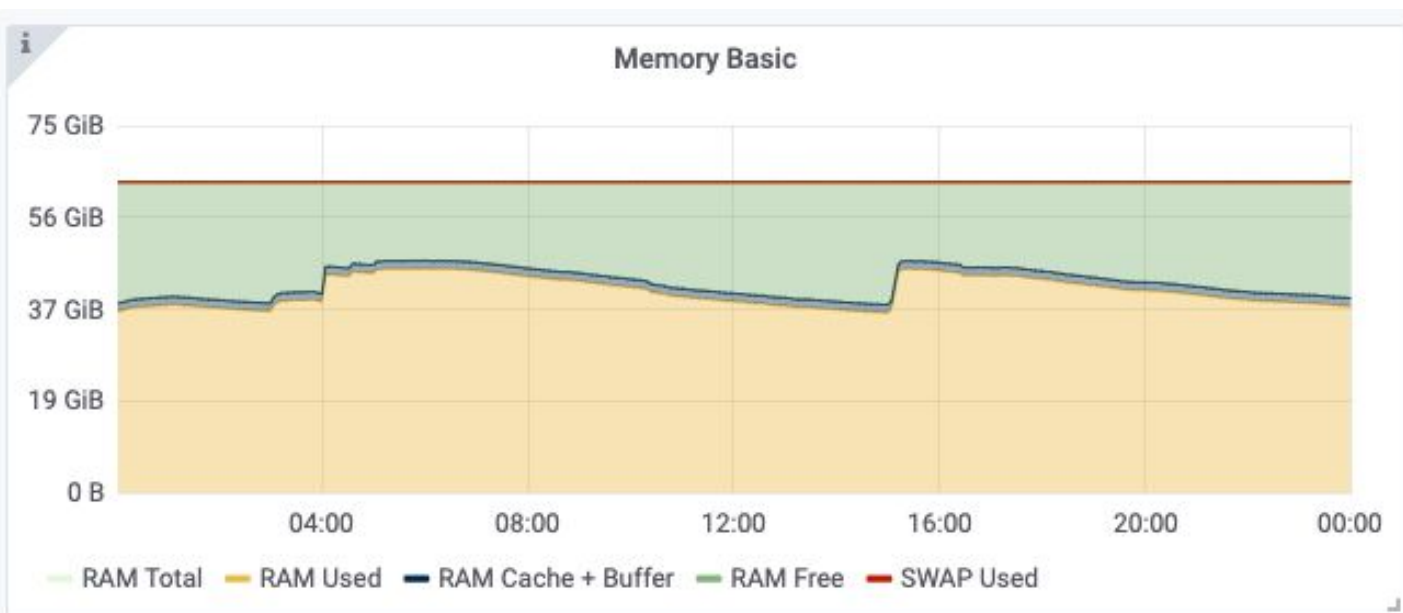
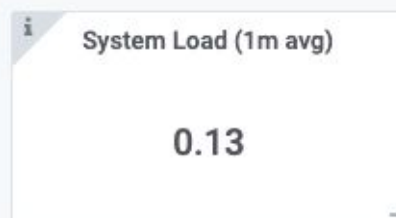
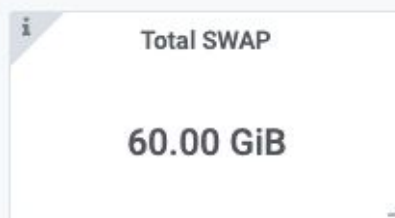
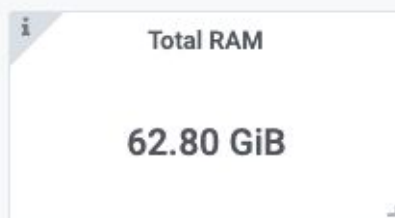
```
ARC Total accesses: 4.57G
Cache Hit Ratio: 94.63% 4.32G
Cache Miss Ratio: 5.37% 245.16M
Actual Hit Ratio: 94.62% 4.32G

Data Demand Efficiency: 77.21% 967.49M
Data Prefetch Efficiency: 14.46% 4.98M
```

Basic CPU / Mem / Disk Gauge



Basic CPU / Mem / Disk Info



談談自動重灌

一切都因為懶

一直點很麻煩、條件好多點選

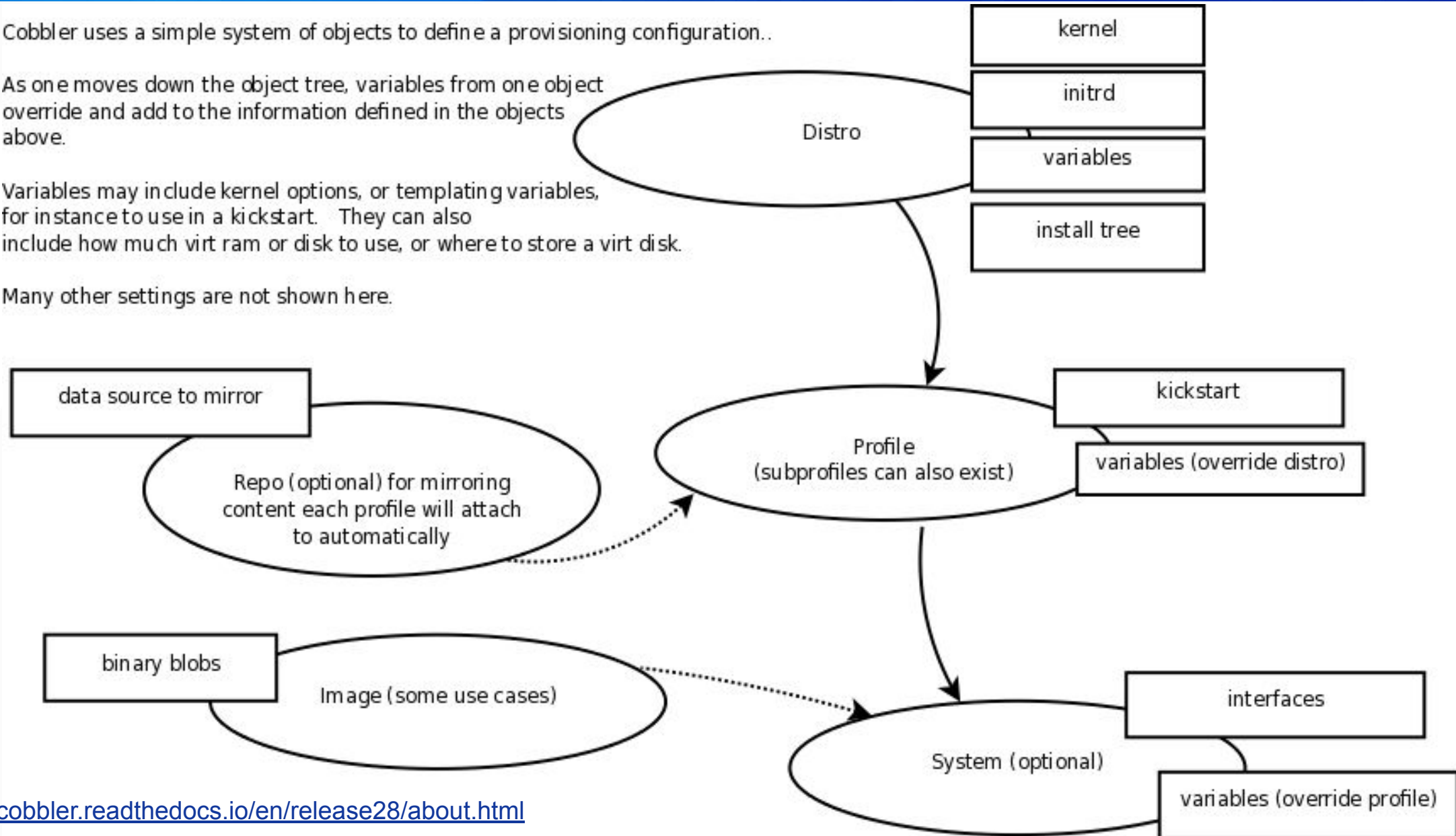
能不能一鍵裝到好？

Cobbler uses a simple system of objects to define a provisioning configuration..

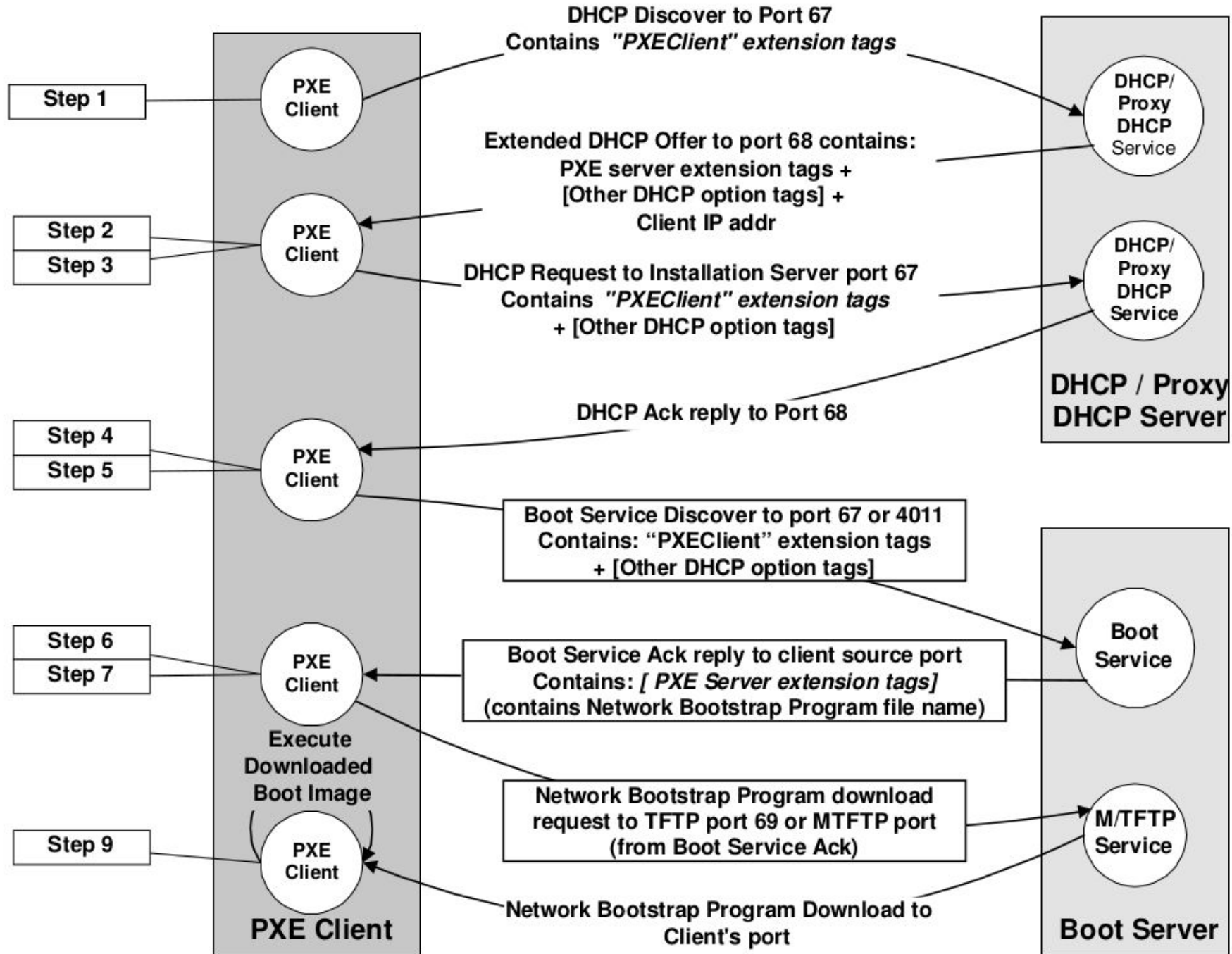
As one moves down the object tree, variables from one object override and add to the information defined in the objects above.

Variables may include kernel options, or templating variables, for instance to use in a kickstart. They can also include how much virt ram or disk to use, or where to store a virt disk.

Many other settings are not shown here.



Ansible 觸發 PXE Client reboot



Cobbler 透過 API 取得各 Server 之參數

準備 Server 參數

- Hostname
- IP, Gateway
- HDD 分割
- 設定 switch vlan

Ansible 觸發工作

- reboot
- 修改 vlan
- cobbler 取得 server 參數

前一頁 PXE 流程

- 硬碟分割
- 安裝套件
(一般安裝 Debian 會做的事)

Ansible 安裝後工作

- 自定套件
- 權限設定
- NIS
- NFS
- 設定環境

```
"tmp-db-0-124": {
  "DNS": [
    "10.1.1.11",
    "10.1.1.12"
  ],
  "gateway": "10.1.1.254",
  "hostname": "tmp-db-0-124",
  "interfaces": {
    "ac:1f:6b:79:ab:30": {
      "vlan": [{
        "id": 1,
        "ipv4": [
          "10.1.0.124/23"
        ]
      }
    ]
  }
},
```

```
    "uplink": {
      "switch": "switch-4DD13-2",
      "FPC": -1,
      "speed": 1,
      "port": "0/20"
    }
  },
  "ac:1f:6b:79:ab:31": {
  },
  "IPMI": "10.2.0.124",
  "partition": [{
    "disk": {
      "size": 120,
      "count": 2
    }
  }
},
```



```
+ "power_user": "",
+ "profile": "debian10.0-x86_64",
+ "proxy": "<<inherit>>",
+ "redhat_management_key": "<<inherit>>",
+ "redhat_management_server": "<<inherit>>",
+ "repos_enabled": false,
+ "server": "<<inherit>>",
+ "status": "production",
+ "template_files": {},
+ "template_remote_kickstarts": 0,
+ "uid": "MTU3NDY3NjMwOS45NzgxNTAzMzQuODUxNw",
+ "virt_auto_boot": "<<inherit>>",
+ "virt_cpus": "<<inherit>>",
+ "virt_disk_driver": "<<inherit>>",
+ "virt_file_size": "<<inherit>>",
+ "virt_path": "<<inherit>>",
+ "virt_pxe_boot": 0,
+ "virt_ram": "<<inherit>>",
+ "virt_type": "<<inherit>>"
+}

changed: [tmp-db-2-49 -> localhost]
Monday 25 November 2019 18:05:10 +0800 (0:00:00.470) 0:00:02.041

TASK [ipmi-reboot-to-pxe : ipmitool set boot devices] *****
changed: [tmp-db-2-49 -> localhost]
Monday 25 November 2019 18:05:10 +0800 (0:00:00.718) 0:00:02.760

TASK [ipmi-reboot-to-pxe : impitool power boot] *****
changed: [tmp-db-2-49 -> localhost]
Monday 25 November 2019 18:05:11 +0800 (0:00:00.703) 0:00:03.463
[]
```

```
1.4523103 pci 0000:00:00.0: enabling extended tags
[ 1.452224] pci 0000:00:0a.0: enabling Extended Tags
[ 1.456291] pci 0000:00:1f.0: quirk: [io 0x0000-0x007f] claimed by ICH6 ACPI
/GPIO/TCO
[ 1.456421] pci 0000:00:1f.0: quirk: [io 0x0500-0x053f] claimed by ICH6 GPIO
[ 1.456522] pci 0000:00:1f.0: ICH7 LPC Generic IO decode 2 PIO at 0ca0 (mask
000f)
[ 1.457114] pci 0000:00:01.0: PCI bridge to [bus 0a]
[ 1.457242] pci 0000:00:02.0: PCI bridge to [bus 09]
[ 1.457548] pci 0000:07:00.0: VF(n) BAR0 space: [mem 0xfbd80000-0xfbd9ffff 6
bit] (contains BAR0 for 8 VFs)
[ 1.457698] pci 0000:07:00.0: VF(n) BAR3 space: [mem 0xfbd80000-0xfbd9ffff 6
bit] (contains BAR3 for 8 VFs)
[ 1.458089] pci 0000:07:00.1: VF(n) BAR0 space: [mem 0xfbe60000-0xfbe7ffff 6
bit] (contains BAR0 for 8 VFs)
[ 1.458238] pci 0000:07:00.1: VF(n) BAR3 space: [mem 0xfbe40000-0xfbe5ffff 6
bit] (contains BAR3 for 8 VFs)
[ 1.458467] pci 0000:00:03.0: PCI bridge to [bus 07-08]
[ 1.458734] pci 0000:06:00.0: 32.000 Gb/s available PCIe bandwidth, limited
by 5 GT/s x8 link at 0000:00:07.0 (capable of 63.000 Gb/s with 8 GT/s x8 link)
[ 1.458922] pci 0000:00:07.0: PCI bridge to [bus 06]
[ 1.459057] pci 0000:00:09.0: PCI bridge to [bus 05]
[ 1.459184] pci 0000:00:0a.0: PCI bridge to [bus 04]
[ 1.459319] pci 0000:00:1c.0: PCI bridge to [bus 03]
[ 1.459838] pci 0000:00:1c.5: ASPM: current common clock configuration is br
```

由 PXE 取得 kernel 開機

Q&A

THANK

YOU

台灣最大社群網站

PIXNET 創立於2003年，2006年成立「優像數位媒體科技股份有限公司」，並於2007年加入城邦媒體控股集團。我們是一間以社群為核心的科技公司，旗下主要服務包含：痞客邦、PIXgoods、PIXmarketing、PIXinsight，透過創新的數據應用、多樣化社群服務，實現「Guide to SMART Life」企業核心價值。2018年，PIXNET 推出「全新痞客邦」加速興趣同好彼此凝聚及交流，並持續與產業各界結盟，踏實建構「社群共榮圈」願景。



PIXNET
Guide to SMART Life



連絡方式

Email: cytseng@pixnet.tw
cytseng@gmail.com



個人 Facebook

後續討論、找工作、找實習都歡迎來信詢問

