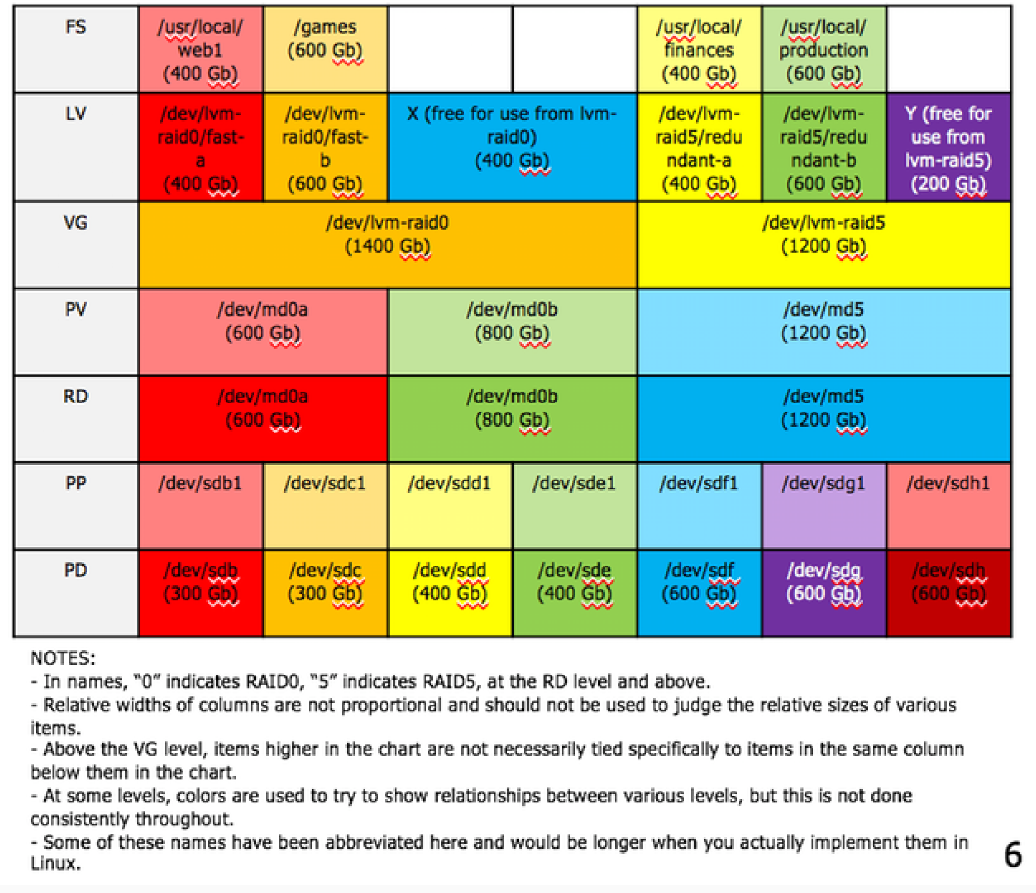
RAID I/O pattern 분석

김서진

LVM on top of RAID

- **Filesystem(FS)**
  - **/usr/local/lvm0**
- **Logical Volume(LV)**
  - **/dev/lvm-raid0/lvm0**
- **Volume Group(VG)**
  - **/dev/lvm-raid0**
- **Physical Volume(PV)**
  - **/dev/md0**
- **RAID device(RD)(via mdadm if Linux SW RAID)**
  - **/dev/md0**
- **Physical Partition(PP)**
  - **/dev/sdb1**
- **Physical Device(PD)**
  - **/dev/sdb**



FS	/usr/local/web1 (400 Gb)	/games (600 Gb)			/usr/local/finances (400 Gb)	/usr/local/production (600 Gb)	
LV	/dev/lvm-raid0/fast-a (400 Gb)	/dev/lvm-raid0/fast-b (600 Gb)	X (free for use from lvm-raid0) (400 Gb)		/dev/lvm-raid5/redundant-a (400 Gb)	/dev/lvm-raid5/redundant-b (600 Gb)	Y (free for use from lvm-raid5) (200 Gb)
VG	/dev/lvm-raid0 (1400 Gb)				/dev/lvm-raid5 (1200 Gb)		
PV	/dev/md0a (600 Gb)		/dev/md0b (800 Gb)		/dev/md5 (1200 Gb)		
RD	/dev/md0a (600 Gb)		/dev/md0b (800 Gb)		/dev/md5 (1200 Gb)		
PP	/dev/sdb1	/dev/sdc1	/dev/sdd1	/dev/sde1	/dev/sdf1	/dev/sdg1	/dev/sdh1
PD	/dev/sdb (300 Gb)	/dev/sdc (300 Gb)	/dev/sdd (400 Gb)	/dev/sde (400 Gb)	/dev/sdf (600 Gb)	/dev/sdg (600 Gb)	/dev/sdh (600 Gb)

NOTES:
- In names, "0" indicates RAID0, "5" indicates RAID5, at the RD level and above.
- Relative widths of columns are not proportional and should not be used to judge the relative sizes of various items.
- Above the VG level, items higher in the chart are not necessarily tied specifically to items in the same column below them in the chart.
- At some levels, colors are used to try to show relationships between various levels, but this is not done consistently throughout.
- Some of these names have been abbreviated here and would be longer when you actually implement them in Linux.

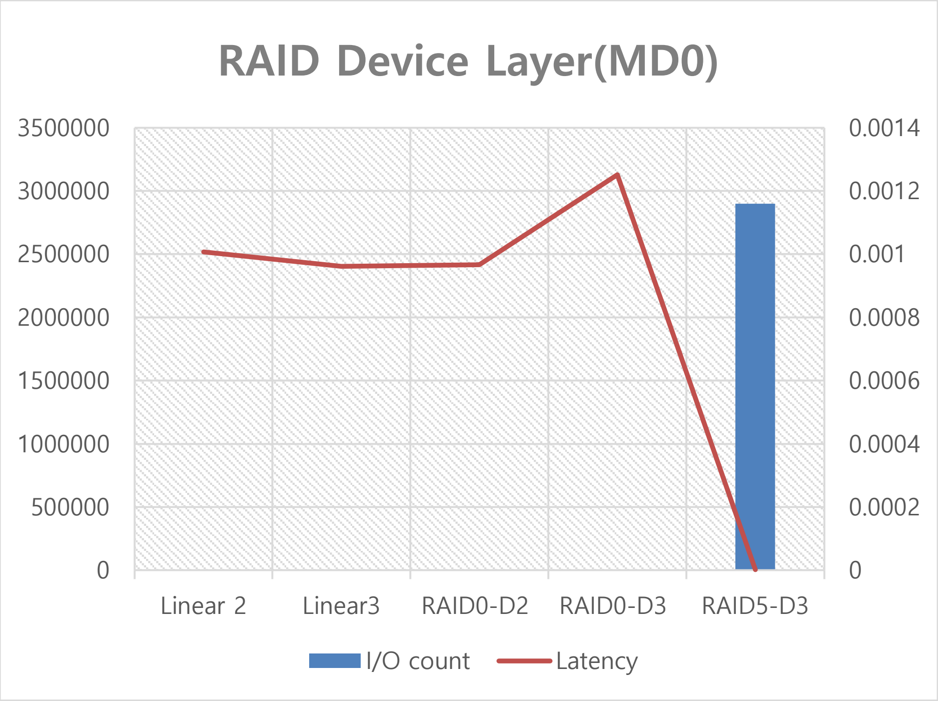
6

실험세팅

- **Fio benchmark v2.1.3**
- **File size: 5GB**
- **Request size:4K , randwrite**
- **File system: ext4 filesystem**
- **Storage:**
  - **Sda: Intel 540s 128GB**
  - **Sdb: Intel 540s 128GB**
  - **Sdd: Intel 540s 256GB**

RAID Device Layer(MD0)

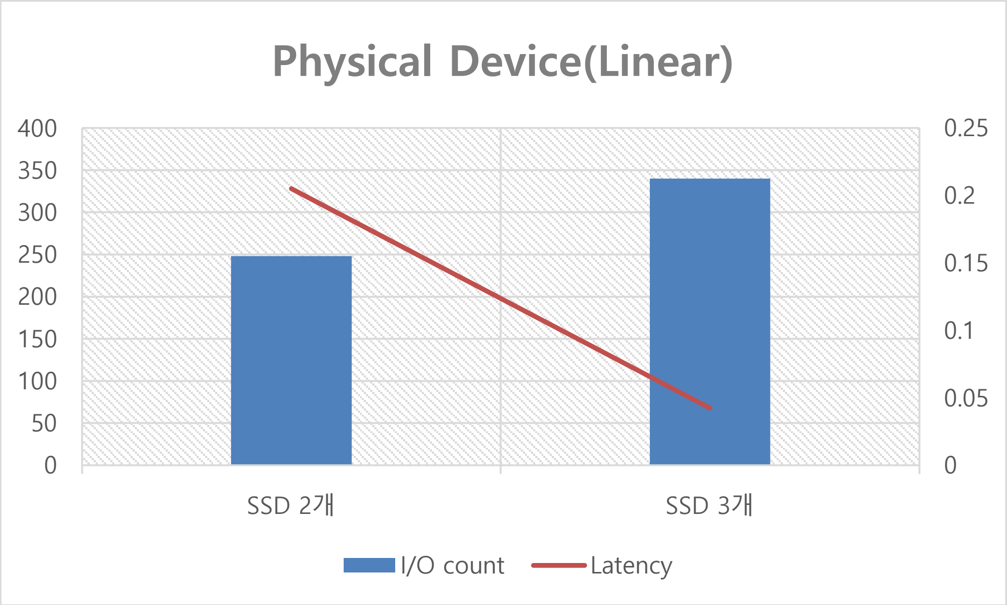
	Linear SSD2	Linear SSD3	RAID0 SSD2	RAID0 SSD3	RAID5 SSD3
I/O count	4	4	4	4	2898241
Latency	0.001007	0.000961	0.000967	0.001251	0.00000158
Write size	8	8	8	8	8



- **RAID Device Layer(MD0)에서**
- **Write size는 8 Block으로 모두 같음에도 불구하고, RAID5의 Latency는 급격히 감소, I/O count는 급격히 증가한다.**

Linear

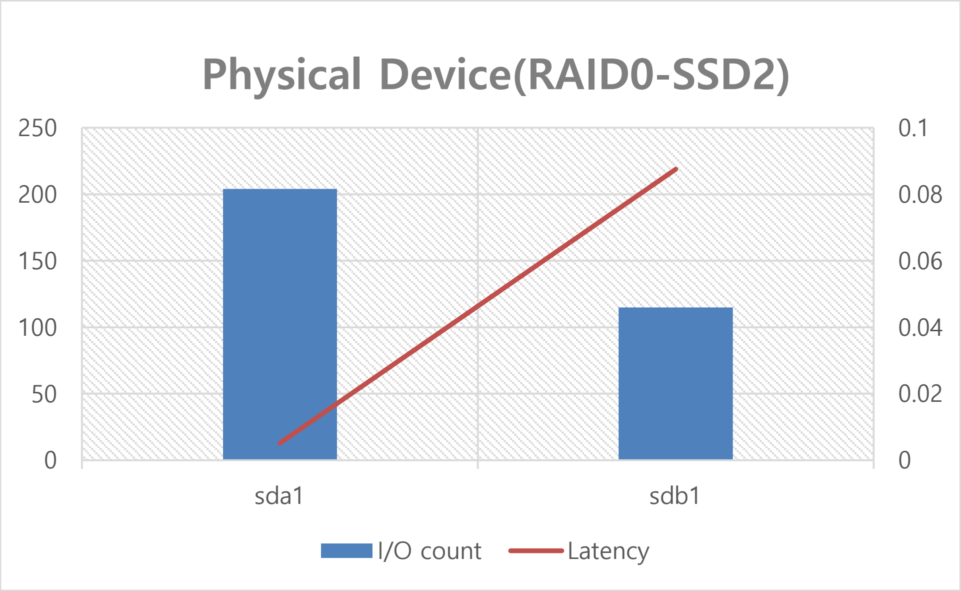
	SSD 2개	SSD 3개
I/O count	248	340
Latency	0.20502	0.042406
write size	13.97736	12.94382



- Linear하게 연결한 SSD array에서, 모두 첫번째 physical device인 sda1에 write 되었지만, I/O count와 Latency는 다르게 나타났다.
- I/O count는 증가했지만, write 되는 size와 Latency는 감소했다.

RAID0, DISK2

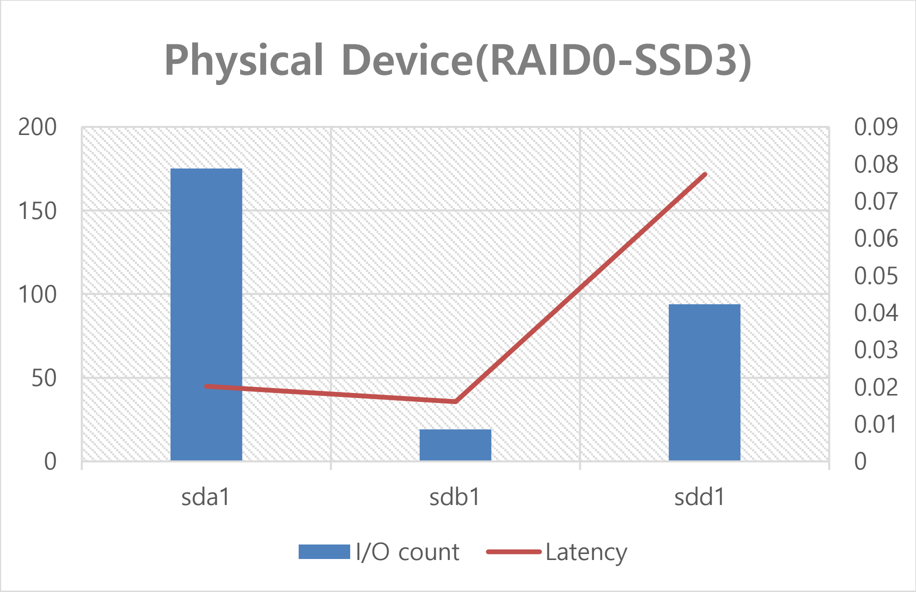
	sda1	sdb1
I/O count	204	115
Latency	0.0051	0.0875
write size	12.9455	14.9565

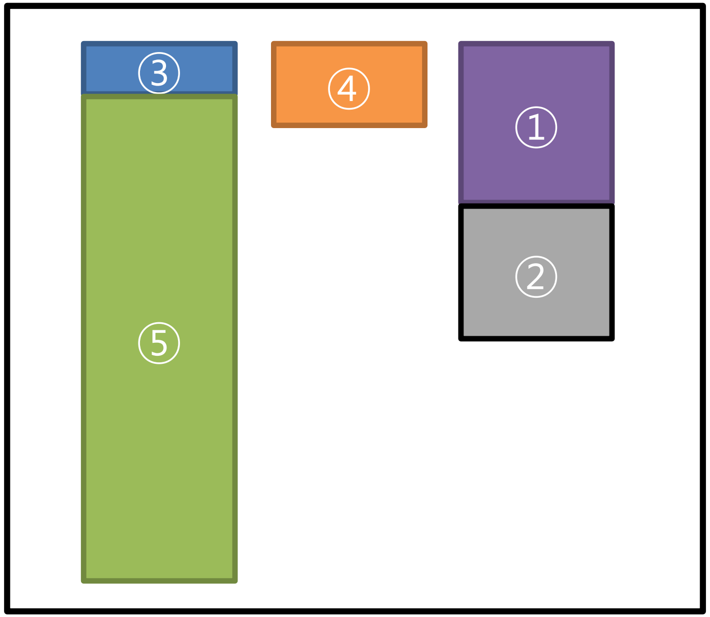


- **Striping 방식을 사용하는 RAID0 에서, 각 Physical device에 write 되는 수는 크게 차이 났다. (2배 가량)**
- **I/O Count는 적어졌지만, Write 의 size는 증가해서 관련이 있을까 생각했지만 (다른 실험으로 확인했을때) 연관이 없는 것으로 나타났다.**

RAID0, DISK3

	sda1	sdb1	sdd1
I/O count	175	19	94
Latency	0.0202	0.0161	0.0772
Write size	15.2686	14.6087	12.9811

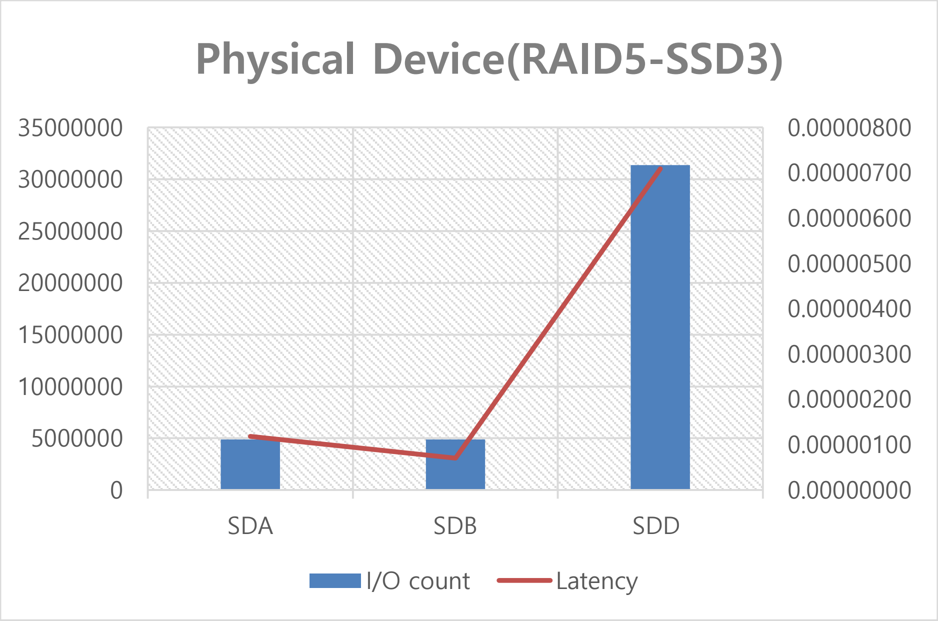




- **RAID0에서 각각 고르게 write 될거란 예상과 달리, device 별로 load balancing이 잘 이루어 지지 않는 것을 확인할 수 있었다.**

RAID5, DISK3

	sda1	sdb1	sdd1
I/O count	4898511	4895736	31373905
Latency	0.00000118	0.00000071	0.00000709
write 크기	13.5300	13.5354	23.8045



- RAID5에서도, 각각의 Physical device별로 I/O count 및 write 되는 양에 큰 격차가 존재했고, latency 또한 마찬가지였다.

Load balancing 원인 및 해결방안

- 원인
  - MDADM tool
  - 각각 SSD 내부 문제
  - SSD 종류
- 줄일 수 있는 방안?
  - Filesystem
  - MDADM code 수정
  - Write 어떻게 되는지

Hardware RAID

- **Boosts system performance for backups and restoration**, (especially in legacy equipment with limited processing power, by adding DRAM cache memory to the system.) Translates to less strain on the system when writing backups, and less downtime when restoring data.
- **Adds RAID configuration options that may otherwise be unavailable using just the motherboard**—like RAID 5/6, for example, which provides one and two drive failure tolerance.
- **Protection against data corruption resulting from a loss of power during the backup process**. Battery backup units (BBU) or onboard Flash memory in RAID cards provide the extra fail-safes here.
- Adds system compatibility with enterprise SAS HDDs, which are designed for 24/7 operation and have extra error correcting features compared to consumer-grade SATA III HDDs.

Software RAID

- 최신식 multi-core server CPU는 많은 system 제약 없이도 백업 및 복구 작업을 실행할 수 있을 만큼 강력하다. Hardware RAID를 사용함으로써 얻을 수 있는 백업 및 복구 성능 향상이 매우 미미할 때.
- Software RAID is used exclusively **in large systems** (mainframes, Solaris RISC, Itanium, SAN systems) found **in enterprise computing.**
- SMBs using NAS devices for backup and restore purposes will find many software-RAID based options: Netgear ReadyNAS; Synology DiskStation(DS), Buffalo TeraStation, are examples.
- Open source인 Software RAID를 사용하는 경우, Flexibility를 얻을 수 있다. Ex. Linux Mdadm
- Software RAID 지원가능한 Hypervisors를 사용하는 경우   ex. Hyper-V -> Storage Space 사용

Hardware RAID and SSD arrays

- System administrators have reported inconsistent performance for certain hardware RAID setups that use flash storage (SSD) arrays. Older RAID controllers disable the built-in fast caching functionality of the SSD that needed for efficient programming and erasing onto the drive. Most current generation RAID controllers give users the option of re-enabling SSD disk caching to alleviate this.
- Having an all-flash storage array set up for RAID 5 provides substantial performance gains compared to a HDD array.

HW RAID SW RAID

- The type of RAID best suits data backup needs will vary from system to system. Hardware RAID is more common in Windows Server environments, wherein its advantages are better realized. Software RAID is more prevalent in open source server systems, wherein its flexibility and comparative low cost of entry make it an attractive option. Both options are completely viable; answering the hardware RAID vs Software RAID question depends on assessing the IT infrastructure—the sever hardware and system administrators operating it—to determine what makes the most sense for any organization.