

Bazy danych a system operacyjny (w Wirtualnej Polsce)

Remigiusz Sokołowski
Wirtualna Polska SA

- Wybór platformy sprzętowej
- **Instalacja i aktualizacja**
- Zarządzanie strukturami bazy danych
- **Archiwizacja i backup**
- **Monitorowanie**
- **Konfiguracja i optymalizacja**

- Konwencje
 - koncepcje
 - Convention Over Configuration (Java)
 - Optimal Flexible Architecture (Oracle)
 - skrypty
 - biblioteki funkcji, (np. /usr/local/lib)
 - centralne repozytorium skryptów (np./usr/local/bin)
 - /lib/lsh (Linux), /lib/svc (Solaris)
 - efekty
 - powtarzalność procesu
 - łatwość wdrożenia nowych członków zespołu
 - szybka automatyzacja pracy
 - efekt synergii - więcej czasu na rozwój technologii

- Narzędzia
 - Menadżery pakietów
 - rpm
 - pkg
 - Systemy zarządzania konfiguracją
 - cfengine (www.cfengine.org)
 - /etc/init.d
 - Solaris Maintenance Framework (Solaris)
 - zasoby /lib/svc
 - zmienne i właściwości SMF (SMF_FMRI, properties)
 - polecenia svcs, svcadm
 - ctid

- Podsystemy IO
 - starsze systemy plików
 - ext3, XFS, UFS
 - journaling
 - wsparcie dla dużych plików (np. opcja largefiles)
 - ZFS
 - autodiagnostyka
 - kompresja
 - zmiana rozmiaru on-line
 - migawki (snapshots)
 - raw devices
 - nowe możliwości – Oracle ASM

- Wirtualizacja
 - implementacje
 - Zony (Solaris)
 - wady i zalety wirtualizacji
 - problemy z użyciem pewnych nastaw
 - podział odpowiedzialności
 - zona + migawki ZFS
 - nakładanie zmian z możliwością wycofania
 - testowanie konfiguracji
 - aktualizacja oprogramowania

- “Zimny” backup/archiwizacja
 - polecenia systemowe
 - cp, scp
 - polecenia systemu backup-owego
 - np. Legato Networker i polecenie save
 - ZFS i migawki (snapshots)

- “Gorący” backup/archiwizacja
 - dlaczego nie polecenia systemowe?
 - backup zarządzany przez użytkownika
 - mysqldump
 - narzędzia specjalistyczne
 - RMAN
 - ibbackup
 - migawki ZFS
 - przy możliwości zablokowania bazy tylko-do-odczytu lub w trybie backup

- CPU
 - uptime
 - vmstat
 - sar [-q]
 - mpstat

```
bash> vmstat 1
```

kthr			memory		page						disk				faults		cpu				
r	b	w	swap	free	re	mf	pi	po	fr	de	sr	f0	m0	m1	m2	in	sy	cs	us	sy	id
0	0	0	2394244	507764	34	107	28	31	31	0	0	0	0	0	0	39629	77840	11122	26	16	58
0	0	0	2549948	497320	0	62	0	0	0	0	0	0	0	0	0	68355	40682	15491	33	22	44
0	0	0	2548784	496044	0	0	0	91	91	0	0	0	0	0	0	62912	48646	15352	25	20	55
0	0	0	2549220	496496	0	0	0	142	142	0	0	0	0	0	0	54956	68490	13959	40	18	42
0	0	0	2549744	497104	0	0	0	0	0	0	0	0	0	0	0	53873	30349	15162	19	17	64
0	0	0	2549272	496632	0	0	0	56	56	0	0	0	0	0	0	30151	32402	11583	14	11	75
0	0	0	2549092	496452	0	0	0	55	55	0	0	0	0	0	0	57843	37749	16743	52	20	28
7	0	0	2548712	496072	0	0	0	16	16	0	0	0	0	0	0	65122	43796	17456	49	26	25

- Pamięć
 - vmstat -p
 - mdb -k (:::memstat)
 - swap [-s | -l], prtswap

- Systemy plików
 - iostat [-xnzC]
 - zpool iostat
 - sar -d
 - fsstat
- Sieć
 - netstat
 - snoop, tcpdump, ethereal

- Aplikacja i jej procesy
 - SMF, skrypty
 - ptools
 - ps, pstree, top (Linux)
 - ps, ptree, prstat, pldd, pstack, pmap, pfiles, etc (Solaris)
 - wewnętrzne mechanizmy baz danych

DTrace

```
#!/usr/sbin/dtrace -qs

io:::start
{
    @bytes[execname, args[0]->b_flags & B_PHYS ? "PHYSICAL"
        : "LOGICAL"] = sum(args[0]->b_bcount);
}

tick-$1
{
    printf("Suma read+write w bajtach\n");
    printa(@bytes);
    exit(0);
}
```

- Konwencje
 - Convention over configuration (Java)
 - Optimal Flexible Architecture (Oracle)
- Narzędzia
 - Menadżery pakietów
 - Systemy zarządzania konfiguracją

- Nastawy systemowe
 - Linux
 - /etc/sysctl.conf
 - /proc/sys (np. echo "8192" >/proc/sys/fs/file-max)
 - sysctl [-w] <parametr>
 - Solaris
 - /etc/system
 - /etc/project

- Istotne parametry
 - Linux
 - `vm.nr_hugepages` – ilość dużych stron pamięci (głównie 2.6)
 - `kernel.shmmax` – max rozmiar pamięci możliwy do zaalokowania dla procesu
 - Solaris
 - `shmsys:shminfo_shmmax`
 - `process.max-file-descriptor`

- Konfiguracja IO
 - Sector zoning
 - poprzez odpowiedni podział na partycje
 - użycie ZFS
 - pisanie do najbardziej “zewnątrznych” bloków
 - Techniki RAID
 - S.A.M.E. (RAID 10)
 - wielkość stripe'ów
 - $\text{db_block_size} * \text{db_file_multiblock_read_count}$
 - 1 MB
 - IO multipathing
 - IP multipathing

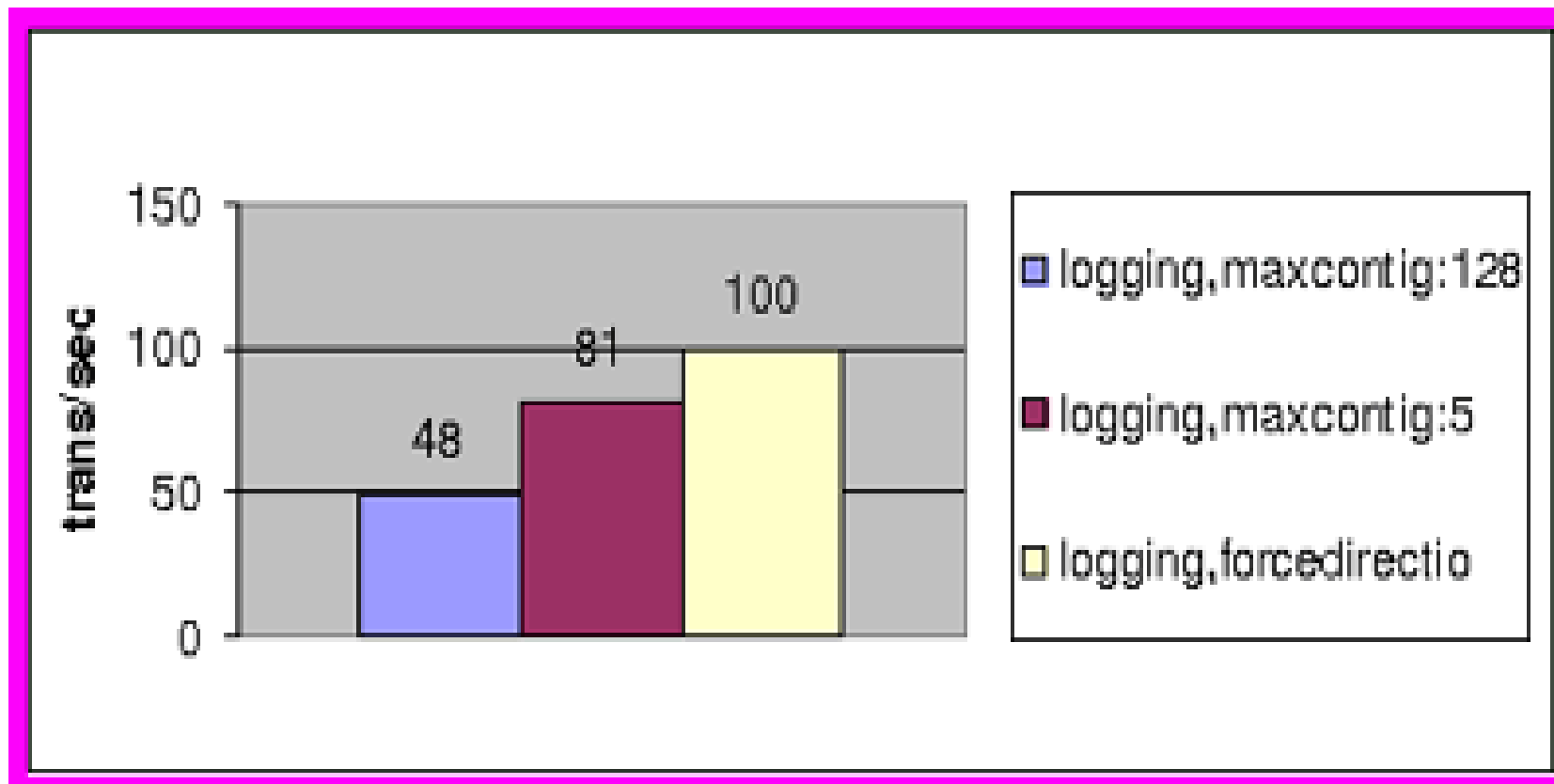
- Zarządzanie procesami
 - menadżery zadań (task schedulers)
 - RT (real time)
 - SYS (ale o nim nie mówimy)
 - TS (time share) - domyślny menadżer
 - FX (fixed priority)
 - FSS (fair share scheduler)
 - # priocntl
 - kontrola wydzielania procesów
 - procesy blokujące krytyczne obszary pamięci (poprzez “zatrzaski”) mogą “zasugerować” przydzielenie większej ilości czasu procesora

- Optymalizacje zarządzania pamięcią
 - biblioteki alokacji (zmienne LD_PRELOAD)
 - umem, mtmalloc, hoard
 - ISM i DISM
 - współdzielące mapowanie pamięci fizycznej na wirtualną
 - duże strony pamięci
 - ważne dla systemów 64-bit i dużych keszy danych
 - wielkość do 4MB (ISM od 2.6, DISM od 9)
 - Intimate Shared Memory (ISM)
 - zablokowane w pamięci fizycznej
 - nie wymagają swap-a
 - Dynamic ISM (DISM)
 - pozwala na dynamiczną rekonfigurację sprzętu

- Opcje pracy systemu plików
 - opcja noatime
 - ograniczenie lub wyłączenie read ahead
 - maxcontig
 - włączenie direct IO
 - opcja forcedirectio
 - direct IO per deskryptor pliku
 - » flaga O_DIRECT, funkcje madvise, posix_fadvise (Linux)
 - » funkcja directio (Solaris)
 - wyłączenie mechanizmu “single writer lock” (Solaris)
 - ok. 90% wydajności “raw devices”
 - asynchroniczne I/O
 - KAIO
 - LWP/threads (Solaris)
 - opcje sync/async (Linux)

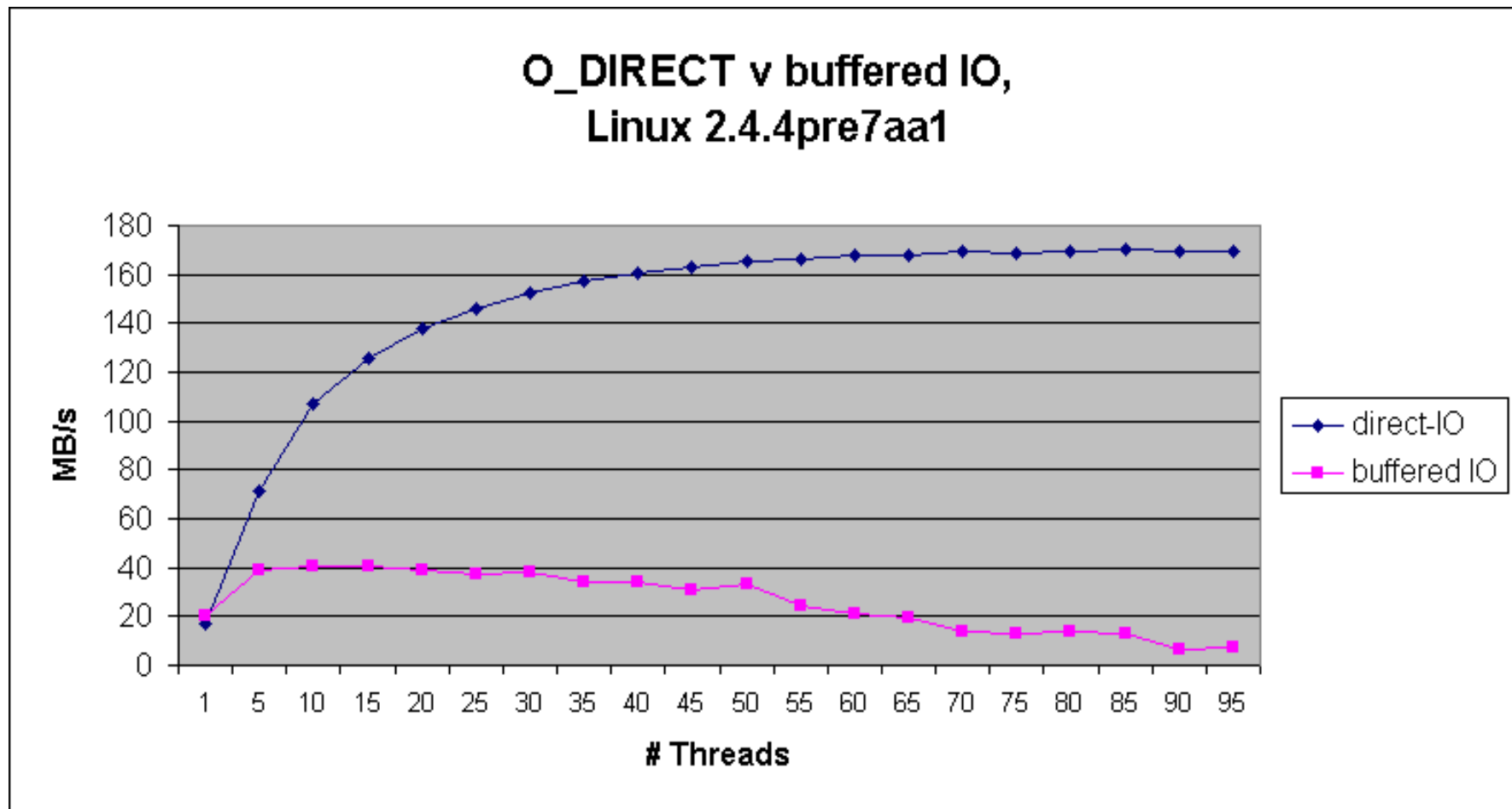
- Opcje pracy systemu plików
 - opcje dla ZFS
 - aktualny ZFS
 - rozdzielenie danych i logów
 - recordsize = db block size (dla plików danych)
 - recordsize = 128k (dla logów)
 - ew. tuning przy pomocy `ztune.sh`
 - `vdev_cache` (6437054)
 - `vq_max_pending` (6457709)

Ilość transakcji na sekundę dla silnika InnoDB



Źródło: http://developers.sun.com/solaris/articles/mysql_perf_tune.html#2

Przykładowa wydajność aplikacji samodzielnie zarządzającej buforem danych



Źródło (wtórne): http://www.ukuug.org/events/linux2001/papers/html/AArcangeli-o_direct.html

- Optymalizacja podstawowych nastaw bazy danych
 - wielkość wewnętrznych buforów danych
 - log powtórzeń
 - wielkość bufora
 - ilość grup
 - metody pracy z podsystemem IO
 - innodb_flush_method (InnoDB)
 - innodb_flush_log_at_trx_commit (InnoDB)
 - filesystemio_options (Oracle)

- Konwencja (choćby i własna)
- Dużo pamięci, ... ale nie za dużo
- S.A.M.E. (1MB \geq szerokość stripe'a \geq wielkość bloku bazy danych)
- Użycie direct IO, raw devices lub ASM dla systemów z własnym cachem, ale ...

- R. McDougall, J.Mauro, “Solaris Internals”
- R. McDougall, J.Mauro, B.Gregg, “Solaris monitoring and tuning”
- A. N. Packer, “ Configuring & Tuning Databases on the Solaris Platform”
- SysAdmin – artykuły
- developers.sun.com
- kerneltrap.org
- www.solarisinternals.com

Trochę linków

http://blogs.sun.com/glennf/entry/where_do_you_cache_oracle

http://www.oracle.com/technology/deploy/availability/pdf/oow2000_same.pdf

http://www.solarisinternals.com/wiki/index.php/ZFS_Best_Practices_Guide

http://blogs.sun.com/realneel/entry/zfs_and_databases