trader media east

A modern HTTP accelerator for content providers

Leszek Urbański
Trader Media East Competence Center

PLNOG 4 – Warsaw, 2010-03-05

PLNOG

trader media east

Varnish

- a state-of-the-art reverse proxy and cache
- open source, initially developed for a Norwegian tabloid “Verdens Gang” in 2006
- Poul-Henning Kamp – architect and lead developer
- Linpro AS

PLNOG

trader media east

Varnish

- used by TOP100 sites
  - Twitter
  - Photobucket
  - weather.com
  - answers.com
  - Hulu
  - Wikia
  - source: Ingvar Hagelund
    http://users.linpro.no/ingvar/varnish/stats-2010-01-18.txt

PLNOG

trader media east

Architecture

- Varnish does not fight the OS kernel!
- uses virtual memory, two main stevedores:
  - mmap()
  - malloc()
- scales well in SMP environments
  - event-based acceptor
  - multi-threaded worker model
- static workspaces – reused, operating on pointers

PLNOG

trader media east

Architecture

- SHM logging
  - an mmap()ed file shared by all threads and logging programs
  - logging without syscalls!

```
memcpy(p + SHMLOG_DATA, t.b, l);
/* or */
vsnprintf((char *)(p + SHMLOG_DATA), mlen + 1, fmt, ap);
```

PLNOG

trader media east

Architecture

- efficient object purging - “ban list”
  - need to purge 200,000 objects from the cache without overloading the server?

```
purge req.http.host == foobar.com && req.url ~ ^/directory/.*$
purge obj.http.Cookie ~ example=true
```

  - Varnish keeps a list of purges
  - every object is tested against the list, but only if requested by a client
  - if it matches, it is refreshed from a backend

PLNOG

trader media east

Architecture

- results?
  - microsecond-level response for cached objects
  - good even for static content
  - performance limit currently unknown :-)
  - **75,000** reqs/s achieved at TMECC
  - **143,000** reqs/s achieved by Kristian from Redpill-Linpro

PLNOG

trader media east

Architecture

- serving a request from cache:

```
<... futex resumed> )           = 0 <0.629910>
futex(0x7f2a577fe2e8, FUTEX_WAKE_PRIVATE, 1) = 0 <0.000011>
ioctl(9, FIONBIO, [0])          = 0 <0.000011>
read(9, "GET /logo.png HTTP/1.0\r\n (...) 8191) = 177 <0.000016>
clock_gettime(CLOCK_REALTIME, {1265632945, 828835974}) = 0 <0.000011>
clock_gettime(CLOCK_REALTIME, {1265632945, 828986444}) = 0 <0.000011>
clock_gettime(CLOCK_REALTIME, {1265632945, 829032564}) = 0 <0.000010>
writev(9, [{"HTTP/1.1"..., 8}, (...) 12912}], 32) = 13227 <0.000039>
clock_gettime(CLOCK_REALTIME, {1265632945, 830411262}) = 0 <0.000011>
close(9)                        = 0 <0.000019>
futex(0x44884bf4, FUTEX_WAIT_PRIVATE, 239, NULL <unfinished ...>
```

- 10 system calls, 4 for clock

PLNOG

trader media east

Features

- run-time management and reconfiguration

```
$ telnet localhost 6082
Trying 127.0.0.1...
Connected to localhost.
Escape character is '^]'.
vcl.list
200 23
active          7 boot

vcl.load new1 /etc/varnish/default.vcl
200 13
VCL compiled.
vcl.use new1
200 0
```

PLNOG

trader media east

Features

- comprehensive logging and management
  - varnishlog
  - varnishncsa
  - varnishtop
  - varnishstat
  - ...and more

PLNOG

trader media east

Features

- logging examples
  - tags

```
varnishtop -i RxURL

varnishtop -i TxURL

varnishtop -i RxHeader -I '^User-Agent'

varnishlog -c -o ReqStart 10.0.0.1

varnishlog -b -o TxHeader '^X-Forwarded-For: .*10.0.0.1'
```

PLNOG

trader media east

Features

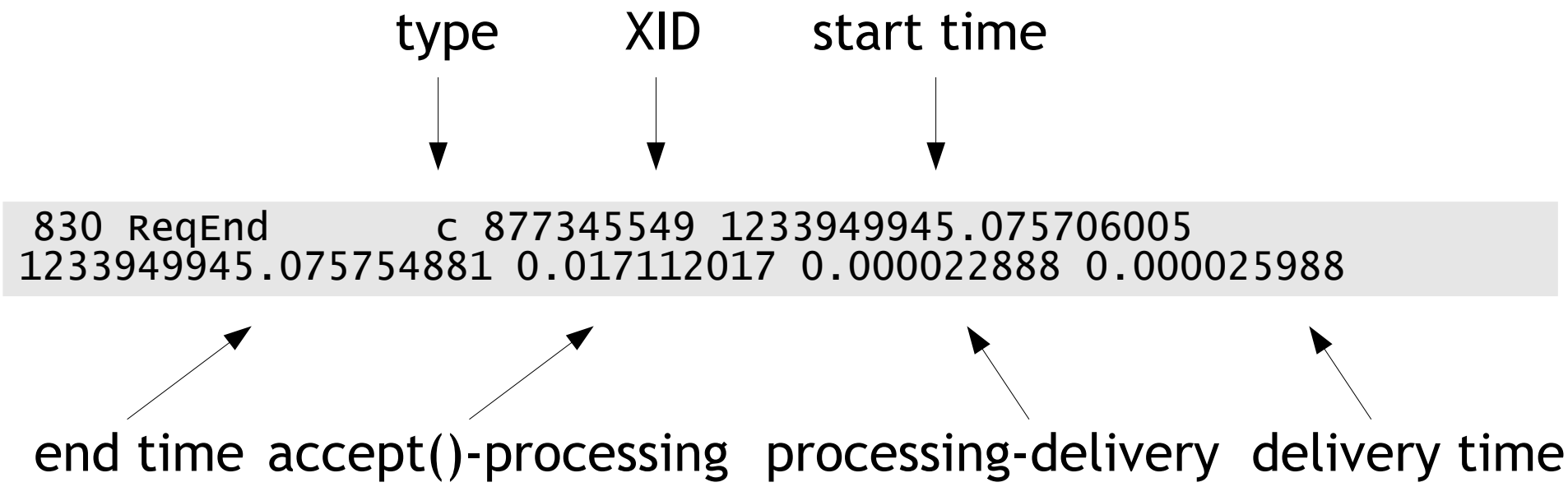
- varnishstat – real time statistics

```
client_conn          87737603         99.74 Client connections accepted
client_req          335496200        381.40 Client requests received
cache_hit           307936704        350.07 Cache hits
cache_hitpass          811746          0.92 Cache hits for pass
backend_conn         12311926         14.00 Backend conn. success
n_object               549675             .   N struct object
n_wrk                     100             .   N worker threads
n_expired            23826372             .   N expired objects
n_lru_nuked                 0             .   N LRU nuked objects
n_wrk_failed                0          0.00 N worker threads not created
s_req               335510357        381.41 Total Requests
s_pass                2947900          3.35 Total pass
s_fetch              27317481         31.05 Total fetch
sma_nbytes         6661407561             .   SMA outstanding bytes
sma_balloc      2173616292374             .   SMA bytes allocated
sma_bfree       2166954884813             .   SMA bytes free
backend_req          27318738         31.06 Backend requests made
esi_parse                   0          0.00 Objects ESI parsed (unlock)
esi_errors                  0          0.00 ESI parse errors (unlock)
```

PLNOG

Features

- timing information



trader media east

Features

- backend load balancing – directors
  - round-robin
  - random
  - client (trunk)
  - URL hashing (trunk)
- grace
- URL serialization
- IPv6 support

PLNOG

trader media east

Features

- Edge Side Includes
  - a markup language for dynamic content assembly
  - also used by Akamai, IBM WebSphere, F5
  - without ESI: page-level caching decisions
  - with ESI: a page can be split into separate blocks and assembled by the cache server

PLNOG

trader media east

Features

- VCL – Varnish Configuration Language
  - a domain-specific language
  - translated to C and compiled
  - dynamically loaded
  - similar to C, Perl
  - = == ! && || ~ !~
  - character escaping like in URLs: %nn
  - no user-definied variables, use HTTP headers:

```
set req.http.something = "";
unset req.http.something;
```

PLNOG

trader media east

Features

- VCL – Varnish Configuration Language
  - “normal” “concatenated” “strings” or

```
{"string
string
"}

synthetic { “string” }
```

  - if () {} elsif {}
  - no loops
  - user defined subroutines
  - regsub(), regsuball()

PLNOG

trader media east

Features

- VCL – Varnish Configuration Language
  - ACLs

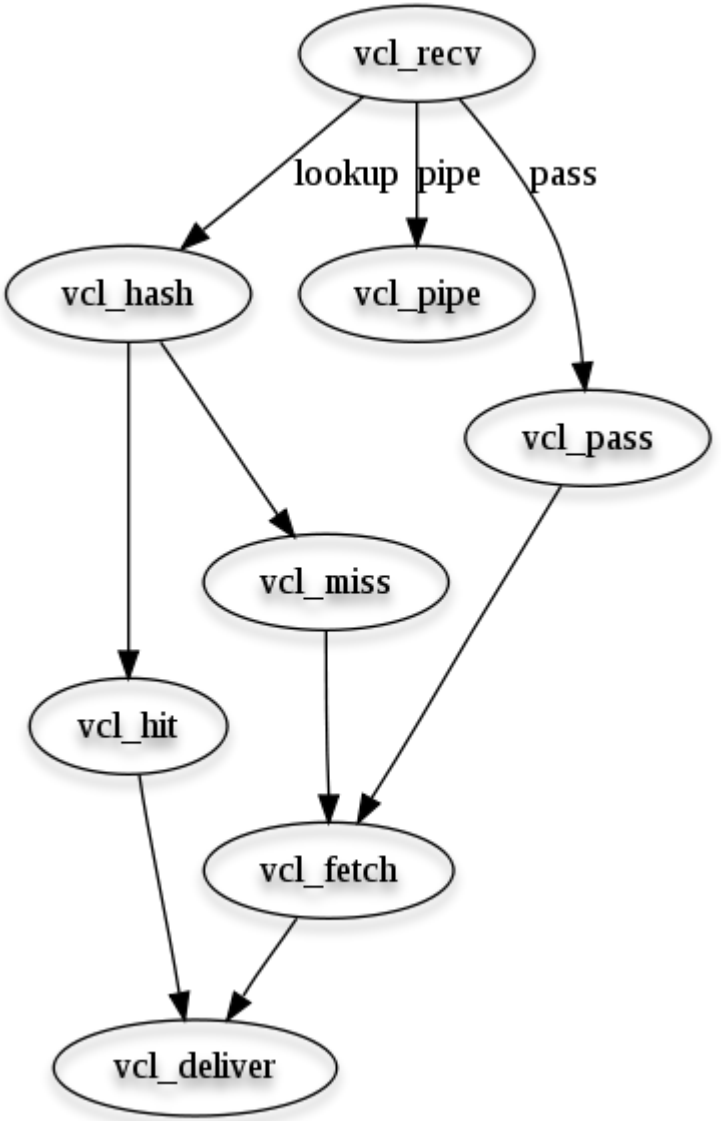
```
acl localnet {
    “localhost”;
    “10.0.0.0/24”;
    ! “10.0.0.1”;
}

if (client.ip ~ localnet) {
    do_magic;
}
```

  - if everything else fails... embedded C!

PLNOG

trader media east

VCL



- request path
  - vcl_recv
  - vcl_pipe
  - vcl_pass
  - vcl_hash
  - vcl_{hit,miss}
  - vcl_fetch
  - vcl_deliver
  - http://varnish-cache.org/wiki/VCLExampleDefault
  - this graph is oversimplified!

PLNOG

trader media east

VCL

- restarts
  - the “restart” keyword turns the request all the way back to vcl_recv, available everywhere

```
sub vcl_fetch {
  if (obj.status >= 500) {
    restart;
  }
}
```

PLNOG

trader media east

VCL

- restarts
  - the “restart” keyword – you can even try another data center

```
sub vcl_recv {
  if (req.restarts == 0) {
    set req.backend = data_center_1;
  } elsif (req.restarts == 1) {
      set req.backend = data_center_2;
  }
}
```

PLNOG

trader media east

VCL examples

- purging, “the squid way”

```
sub vcl_recv {
    if (req.request == "PURGE") {
       if (!client.ip ~ purge) {
         error 405 "Not allowed";
       }
    lookup;
    }
}
sub vcl_hit {
    if (req.request == "PURGE") {
      set obj.ttl = 0s;
      error 200 "Purged";
    }
}
sub vcl_miss {
    if (req.request == "PURGE") {
      error 404 "Not found";
    }
}
```

PLNOG

trader media east

VCL examples

- URL rewriting

```
if (req.http.host ~ "^(www\.)?foo" && req.url ~ "^/images/") {
  set req.http.host = "images.foo";
  set req.url = regsub(req.url, "^/images/", "/");
}
```

- redirects (a bit of a hack)

```
sub vcl_recv {
  if (req.http.host = "^(www\.)?foo.com" && req.http.User-Agent ~
                        "iPhone|Nokia|Motorola") {
    error 701 "Moved temporarily";
  }
}
sub vcl_error {
  if (obj.status == 701) {
    set obj.http.Location = "http://m.foo.com/";
    set obj.status = 302;
    deliver;
  }
}
```

PLNOG

trader media east

VCL examples

- cookie based hashing

```
sub vcl_hash {
    if (req.http.Cookie ~ "language=esperanto" ) {
      set req.hash += "LangEsperanto";
    }
}
```

  - result: a separate cached version of the object for requests with Cookie: language=esperanto;
- extracting the value of a cookie
  - nothing more than a regexp

```
regsub(req.http.Cookie, "^.*?cookie=([^;]*);*.*$", "\1");
```

PLNOG

trader media east

Best practices

- object TTL control – headers from backend (a.k.a. RFC 2616)
  - considered in the following order:
  - **Cache-Control: s-maxage**=<relative time>
  - **Cache-Control: max-age**=<relative time>
  - Varnish **ignores** all other Cache-Control headers (unless told otherwise in VCL)
  - **Expires:** absolute time, requires synced clocks

PLNOG

trader media east

Best practices

- object TTL control – VCL

```
sub vcl_hit {
    if (req.http.host ~ “^images\.”) {
      if (obj.hits > 5 && obj.hits < 10) {
        set obj.ttl = 8h;
      } elsif (obj.hits >= 10) {
         set obj.ttl = 2d;
      }
    }
}
```

- Varnish sets the Age: header

- if in doubt, check varnishlog
  - TTL tag

PLNOG

Best practices

- caching policy
  - Vary
  - Content-Length

trader media east

Best practices

- compression
  - Varnish leaves compression up to the backends
  - gzip, deflate, none – data set * 3
  - normalize Accept-Encoding from browsers

PLNOG

trader media east

Best practices

- sanitize request headers
  - we've had requests from a partner coming in to “http://our.com/?http://another.com/.*”

```
if (req.url ~ "^/\?http://") {
  set req.url = regsub(req.url, "\?http://.*", "");
}
```

  - cache hit ratio went from 92% to 94%

- hit ratio and backend requests: 1% is half of 2%!

PLNOG

trader media east

Best practices

- test
  - wget --save-headers
  - curl -i
  - LWP: GET -USsed
  - caveat: lwp-request does: “GET http://foo/bar”

PLNOG

trader media east

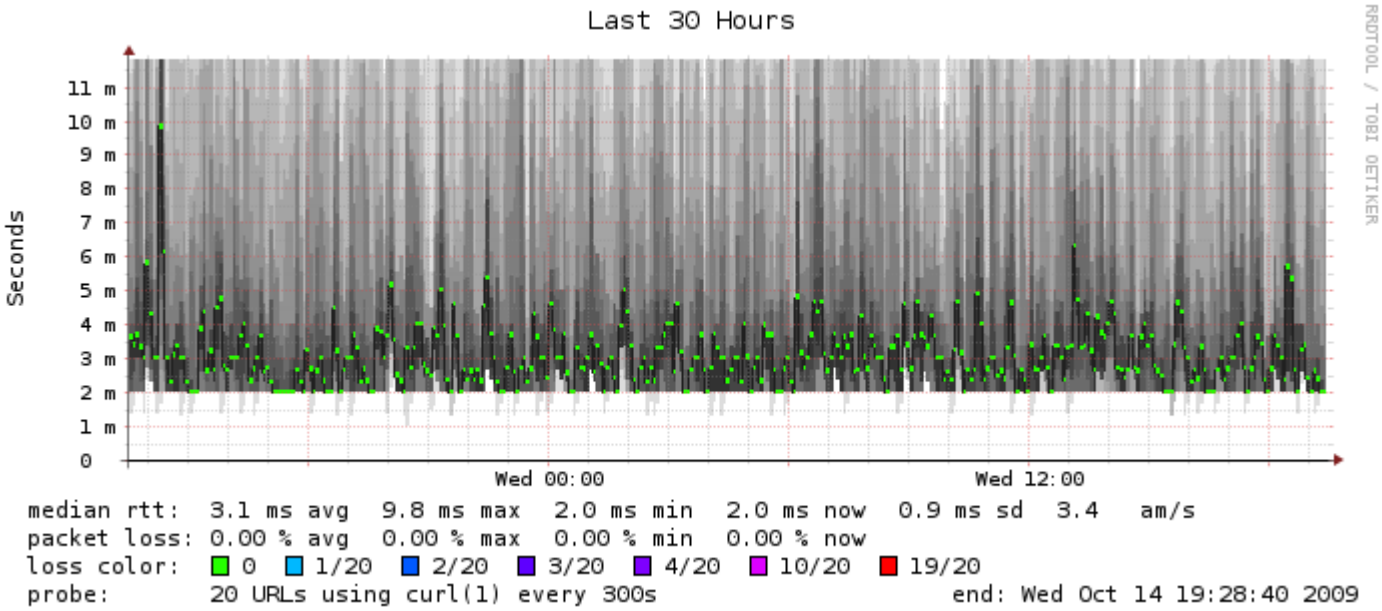
OS environment

- forget about 32-bit
- malloc() better than mmap() for in-memory cache sets
- sometimes better for larger-than-memory cache sets on Linux (YMMV)
- Varnish on virtualized guests?
  - slight latency difference
  - can be an issue for on-line auction sites

PLNOG

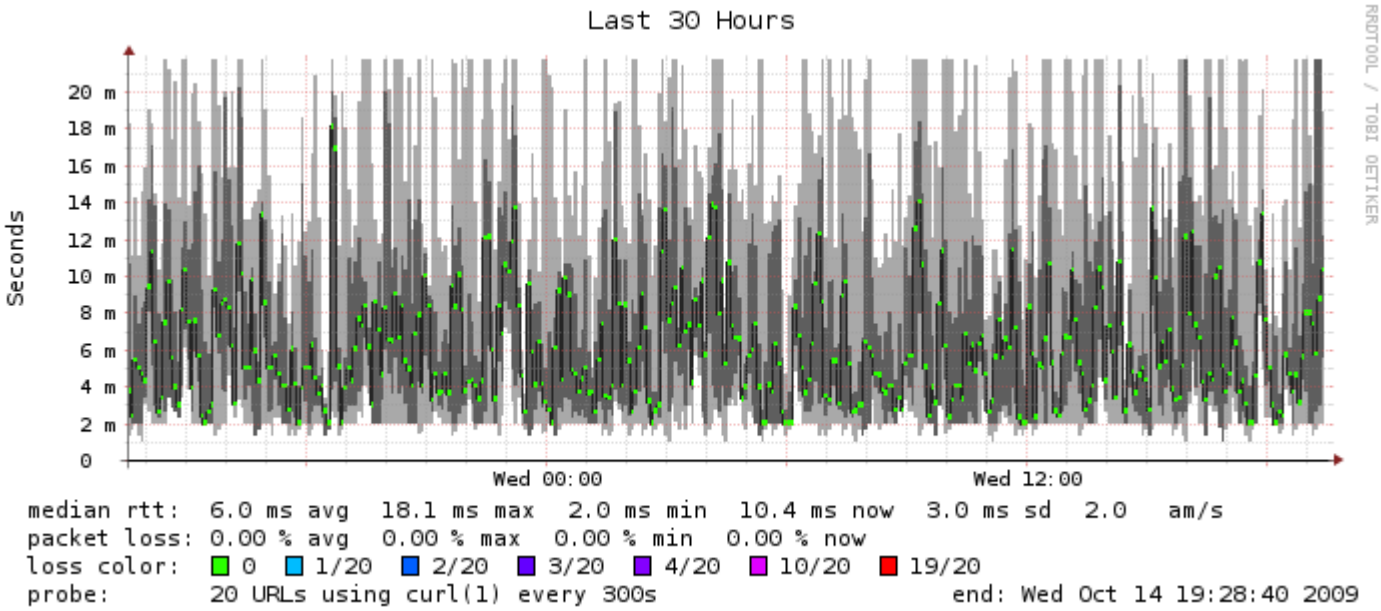
OS environment

- Virtualization
  - Varnish on a standalone system



OS environment

- Virtualization
  - Varnish on a Xen domU with pinned vcpus



trader media east

OS environment

- I/O related tuning on Linux
  - set vm.swappiness to 0
  - **/var/lib/varnish/$HOSTNAME/_.vsl** – the SHM log
  - put the SHM log on tmpfs
  - anticipatory elevator best on HDDs, noop on SSDs
  - use ext2
  - noatime
  - swap striping
  - iSCSI is great for logs

PLNOG

trader media east

OS environment

- network tuning
  - run NTP
  - check if your load balancer uses keep-alive
  - /proc/sys/net – don't tune if you don't know what you're doing

PLNOG

trader media east

Thank you

Questions?

PLNOG

trader media east

Sources

- http://varnish-cache.org/
- TMECC VCL configs
- Wikia VCL configs
- http://kristian.blog.linpro.no/
- http://ingvar.blog.linpro.no/
- #varnish
- unabridged version of this presentation available at: http://slideshare.net/tgr1

PLNOG