NAME

userver − Micro-server
(A fairly simple web server designed for experimentation)

SYNOPSIS

userver

[**-A**|**--accepts-only**] [**-A**|**--close-after-accept**] [**-B**|**--block-for-send_events**] [**-C**|**--caching-on**] [**-D**|**--debug-mask** *0xN*] [**-E**|**--send-events**] [**-F**|**--reset-on-close**] [**-G**|**--use-getpid**] [**-H**|**-h**|**-?**|**--help**] [**-I**|**--rcv-sock-low-wat** *N*] [**-J**|**--snd-sock-low-wat** *N*] [**-K**|**--kdebug-mask** *0xN*] [**-L**|**--full-read** *N*] [**-M**|**--multi-accept**] [**-N**|**--send-loop**] [**-O**|**--select-timeout** *N*] [**-P**|**--count-sigpipes**] [**-Q**|**--ecb-hi-water** *N*] [**-R**|**--read-sockbuf-size** *N*] [**-S**|**--dont-intr-select**] [**-T**|**--port** *N*] [**-U**|**--auto-accept**] [**-W**|**--write-sockbuf-size** *N*] [**-Y**|**--send-io-events**] [**-a**|**--async-mode**] [**-b**|**--ecb-buf-entries** *N*] [**-c**|**--max-conns** *N*] [**-d**|**--delay** *N*] [**-e**|**--extra-accepts** *N*] [**-f**|**--max-fds** *N*] [**-g**|**--accept-on-close**] [**-i**|**--interactive**] [**-j**|**--kernel-info**] [**-k**|**--sigio-use-proc-mask**] [**-l**|**--listenq** *N*] [**-m**|**--accept-count** *N*] [**-n**|**--no-accepts**] [**-n**|**--listen-only**] [**-o**|**--process-fd-order** [**up**|**down**|**writes-down**|**writes-up**|**lru**|**lifo**|**fifo**]] [**-p**|**--procs** *N*] [**-q**|**--ecb-low-water** *N*] [**-r**|**--eager-reads**] [**-s**|**--conns-in-server-loop**] [**-t**|**--free-fd-threshold** *N*] [**-u**|**--track-max-fd** *N*] [**-w**|**--eager-writes**] [**-x**|**--sigio-accepts**] [**-y**|**--memcpy**] [**-z**|**--rejection-rate**] [**--close-after-sock-init**] [**--close-after-read**] [**--close-after-parse**] [**--send-polls-for-accepts**] [**--cache-table-size** *N*] [**--cache-max-bytes** *N*] [**--cache-max-file-size** *N*] [**--cache-max-load-factor** *N*] [**--cache-table-print**] [**--cache-lock-pages**] [**--cache-for-spec**] [**--use-sendfile**] [**--use-accept-send**] [**--use-madvise**] [**--trace-summary**] [**--trace-summary-only**] [**--use-poll**] [**--use-epoll**] [**--use-epoll2**] [**--use-epoll-ctlv**] [**--ip-addr** *N*] [**--stats-interval** *N*]

DESCRIPTION

userver is a micro web server that is meant to be used to experiment with the design and implementation of web servers. In particular the original intention was to permit experimental comparisons of different event dispatch mechanisms within the same appliication framework. For this reason there are lots and lots and lots of parameters. They are mainly used to control the behaviour of the server.

Some of the design philosophy and experimental results obtained using this server are described in [Brecht-2001]. One of note is that all data structures are allocated at initialization and are not resized (grown). This helps to ensure that experiments can be repeated because with dynamic sizing of data structures different loads could result in different server behaviour.

The userver uses an event driven architecture. The idea is to run one event driven server per processor. If using sendfile on Linux the file system buffer cache will be shared across all servers. For high-performance one would want to use sendfile since it doesn't involve extra copying from the file buffer cache to socket buffers (i.e., it's a zero copy implementation).

IMPORTANT: This micro web server is not meant to be a full-blown web server. It is specifically designed for conducting performance experiments. For that reason design decisions were made in favour of producing repeatable experiments.

EXAMPLES

userver This starts the **userver** with it's all of its default options.

userver --max-conns 300 --accept-count 25

This command causes the **userver** server to limit the number of simultaneous connections to 300. When an event indicating that a new connection is requested accept is called continuously until either 25 new connections are accepted, or there are no more outstanding connections.

userver --max-conns 300 --accept-count 0

This command causes the **userver** server to limit the number of simultaneous connections to 300. When an event indicating that a new connection is requested accept is called repeatedly until there are no more outstanding connections. (Note that --accept-count=0 means an infinite number of connections, i.e., no limits).

userver --ip-addr 192.168.10.105 --max-conns 300 --listenq 128

userver --ip-addr 192.168.20.105 --max-conns 300 --listenq 128

This will start up two copies of the userver each accepting connections from a different interface.

userver --ip-addr 192.168.10.105:6801 --max-conns 300 --listenq 128

userver --ip-addr 192.168.20.105:6802 --max-conns 300 --listenq 128

Same as above except one listens on port 6801 and the other on port 6802.

userver --port 6800 --max-conns 300 --listenq 128

This command causes **userver** to listen for incoming connection requests on port 6800. It uses a listen queue of length 128 and it will permit up to 300 simultaneous connections. Once there are 300 open connections it will not accept new connections until an existing connection is closed.

userver --port 6800 --max-conns 300 --listenq 128 --use-poll

This is the same as above except using *select()* calls to get events the **userver** will use *poll()*.

userver --port 6800 --max-conns 300 --listenq 128 --use-epoll

This is the same as above except using *poll()* calls to get events the **userver** will use *epoll()*.

userver -C --cache-table-size=10000 --cache-max-bytes=262144000 --cache-max-file-size=1048576

Caching is turned on. Caching parameters limit the number of files that are cached to 10000, the total number of bytes cached to 250 MB, and the size of the largest file that can be cached to 1 MB.

OPTIONS

The operation of **userver** can be controlled through a number of options. The tool supports both short (one-character) and long (arbitrary-length) option names. Short options are prefixed with a single leading dash (-), long options with a double-dash (--). Multiple short options can be grouped together (e.g., "**-rw**" is equivalent to "**-r -w**") and long options can be abbreviated so long as they remain unique. Parameters to options can be specified either by following the long option name with an equal sign and the parameter value (e.g., **--port=6800**) or by separating the option name and value with whitespace (e.g., **--port 6800**).

OPTIONS (GENERAL)

-c N

--max-conns=*N*

The maximum number of simultaneous connections the server will handle. Note that this is typically limited by the number of open file descriptors permitted. Once the maximum number of open connections is reached incoming connection requests are refused until an existing connection is close. Note that idle connections can be timed out eventually (but this hasn't been tested lately). This is also used to size a number of internal data structures that are allocated at initialization time.

-f N

--max-fds=*N*

The maximum number of open file descriptors permitted. I can't recall if this is actually enforced but it is used to size internal data structures. Note that there is a relationship between **--max-conns** and **--max-fds.** If files are not being cached you could (in the extreme case) have one fd used for the socket connection and one fd used for the open file. So one needs to set **--max-fds** = more than 2 x **--max-conns.** Note: it should be greater because of some descriptors that are used up by stdin, out, err, and a few open files, etc.

-m N

--accept-count=*N*

Call accept repeatedly until either, **--max-conns** is reached, **accept()** returns EWOULDBLOCK, or *N* new connections have been accepted. Note that the value of *N* can make a substantial difference in the performance of event dispatch mechanisms under heavy loads. See [Brecht-2001].

-T N

--port=*N*

Specify the port number that the server should listen for connections on.

-l N

--listenq=*N*

Specify the length of the appliation's listen queue. This specifies how many connection requests are allowed to be queued in the application before the kernel rejects incoming connection requests.

-L[1 | 2]

--full-read=[1 | 2]

Affects how calling read works. If this is set to 1 the server will loop on reading the socket until the read fails. This mainly happens either because there is nothing left to read because the other end has been closed or until it would have blocked (EWOULDBLOCK).

If a value of 1 is used the loop on reads will only occur on calls to read that occur as a result of an event indicating that the first read should be done. If a value of 2 is used the loop on reads occur as above but if **--eager-reads** is also set the server will loop on eager reads (i.e., those that will be done immediately following a the acceptance and setup of a connection).

-H

-h

-?

--help Print out the usage message.

-M N

--multi-accept=*N*

Use the *multiaccept()* system call to perform accepts. NOTE: this was a system call that we added to and experimented with in Linux. The idea here is that since we are repeatedly calling *accept()* to accept a bunch of connections, why not provide a system call that allows up to a maximum of *N* connections to be accepted in one system call.

-S

--dont-intr-select

Disable asynchrnous event notification when calling *select()*. I believe that this was initially meant to work for SIGIO and for SEND.

-d N

--delay=*N*

The idea was to introduce a delay that might simulate work being done by the server (like some computation and/or a dynamic request). I couldn't find a good method for delaying for a specified amount of time (note that it should be small) with a enough accuracy.

-e N

--extra-accepts=*N*

While in the middle of processing events that have been obtained (e.g., from select, or some other mechansim) periodically poll for new connection requests. In this case every time *N* descriptors are processed we check for new connection requests. Note that this open was never used effectively. I found it easier to use and tune the **--accept-count** option.

-g

--accept-on-close

When a connection is closed check to see if there are any outstanding connection requests. Note: that this only occurs if the maximum number of connections was reached and now a closed connection will reduce the number of connections below **--max-conns.**

-o [up|down|writes-down|writes-up|lru|lifo|fifo]

--process-fd-order

[up|down|writes-down|writes-up|lru|lifo|fifo]

Change the order in which file descriptors are handled. They are numbered from 0 to *N* so the uptions are:

up: from 0 to N

down: from N to 0

writes-down:
to writes from N to 0, then reads from N to 0

writes-up:
to writes from 0 to N, then reads from 0 to N

lru: least recently descriptor touched first

lifo: first descriptor (connection) added to the set first

filo: first descriptor (connection) added to the set last

-p N

--procs=*N*

Start *N* copies of the server. Note that this option hasn't been tested in a long time and is almost certain to cause problems.

-s

--conns-in-server-loop

Add an additional poll for new connections each time through the server loop.

-t N

--free-fd-threshold=*N*

When the maximum number of connections **--max-conns** is reached the server stops checking for incoming connections. This options controls when the server starts checking again. It starts checking again after *N* connections can be accomodated before reaching **--max-conns.**

-u

--track-max-fd

The server currently keeps track of the maximum fd ever used and uses that as a parameter to calls like select. This option implements code to dynamically keep track of the maximum file descriptor currently used. Note that this requires a bit of extra processing but it's probably not noticable, especially if it can save estra overhead on copying fdsets and processing time in event mechanisms (e.g., select).

-r

--eager-reads

try to eagerly read from new connections. Call **read() to try to read the request as soon as** the new connection is accepted. In this mode the server optimistically assumes that data will be available for reading when a connection is made on a new socket (or very shortly after). If the assumption is incorrect, the read call simply returns with the error EWOULDBLOCK. Later the event notification mechanism will indicate when the socket is readable.

-w

--eager-writes

try to eagerly perform writes to new connections when the response is available. Both **--eager-read** and **--eager-write** also try to take advantage of any potential locality effects by working on the most recently used file descriptor and socket.

-y

--memcpy

Use the **memcpy** library call to copy interest sets before calling select. This is done rather than doing a straight assignment. This exists because the size of the interest set could be large and I don't know how clever a compiler would be about assigning large data structures. NOTE: it currently copies the side of an fd_set. This should be changed to only copy the amount of data required for the maximum fd of interest.

-z

--rejection-rate

Use the **rejection-rate** is not yet implemented. The ideas would be to specify the frequency with which new connection requests should be rejected.

--use-sendfile

Use the **sendfile()** call on systems that have it available. This is only used for uncached files.

--use-madvise
Use madvise() calls to provide the kernel with hints that files are being read sequentially and when they should no longer be cached.

--use-poll
Use poll() to get events instead of select.

--use-epoll
Use epoll() to get events instead of select.

--use-epoll2
Use epoll() to get events instead of select but attempt some optimizations to reduce the number of epoll_ctl calls. We found the number of epoll_ctl calls to be excessive so this options calls epoll_ctl when a new connection is added and sets the interest to be READ and WRITE. This means that epoll_wait may return events that we aren't interested in (e.g., that a socket is writable even though we are busy reading) but we are exploring the tradeoffs between the two approaches (--use-epoll and --use-epoll2).

--use-epoll-ctlv
Use epoll_ctlv() to update events. This requires the addition of a new system call, epoll_ctlv. This call permits one to collect a bunch of epoll_ctl calls and change interest in a bunch of fds at once instead of having to do one system call (epoll_ctl) per change.

--ip-addr N

--ip-addr N:P
Specify an IP address to listen for new connections on. Optionally takes a port number P as well. The special option --ip-addr 0.0.0.0 specifies that the userver should listen on any address (this is the default).

--stats-interval N
Print out some simple stats about related to server performance at a regularly specified interval. The interval is specified in seconds. NOTE: this options overrides the --select-timeout option in order to be able to print out stats according to the selected interval.

OPTIONS FOR CONTROLLING SOCKETS

-F

--reset-on-close
Calls *shutdown()* instead of *close()*.
WARNING do not use this for a real system!!!

-I N

--rcv-sock-low-wat=*N*
Sets the sockets receive buffer's low water mark to *N*. Not available on all systems.

-JN

--snd-sock-low-wat=N
Sets the sockets send buffer's low water mark to *N*. Not available on all systems.

-O N

--select-timeout=*N*

Set the timeout option to the *select()* system call.

-R N

--read-sockbuf-size=*N*

-W N

--write-sockbuf-size=*N*

Set the size of the read or write socket buffers. Note that increasing these may help significantly for large transfers.

--use_tcp_cork

Cork the TCP queue when writing the header for the reply. Currently only implemented for the sendfile portion of the code.

OPTIONS FOR USE WITH SIGIO

-a

--sigio-accepts

Use asynchronous notification via SIGIO to accept new connections. Asynchronous notification may sound costly at first but consider that some form of asynchronous notificaion may be the only way for an application that is in the middle of something (e.g., computation) to find out about a connection requests.

-k

--sigio-use-procmask

To disable and later enble asynchronous notification of incoming requests when using SIGIO one can use **fcntl()** to turn off and on asynchronous notification or use **sigprocmask().** Be careful of race conditions. By default the server uses **fcntl()** but this option says to use **sigprocmask()** instead.

OPTIONS FOR CACHING

-C

--caching-on

Turns caching on.

--cache-table-size=*N*

Specifies the size of the hash table used to cached files. One entry is used per cached file. NOTE: if the **--cache-max-load-factor** is exceeded the server will print a message and exit. We do not automatically increase the size of the hash table because we want to ensure identical and repeatable behaviour from one experiment to the next.

--cache-max-load-factor=*N*

Specifies the maximum number of entries that can be filled in the hash table, expressed as a percentage of the table size. For example **--cache-max-load-factor = 70** will not permit the hash table to become greater than 70% full. If it does, a message is printed and the server exits.

--cache-max-bytes=*N*

Specifies the maximum number of bytes cached.

--cache-max-file-size=*N*

Don't cache files larger than *N* bytes.

--cache-table-print

When the program ends, print out the contents of the hash table used for caching.

--cache-for-spec

This enables a special hash function that does perfect hashing on URLs requested by SpecWeb99. This is intended for experimentation only.

--cache-lock-pages

Lock all cached pages into memory (must run as root for this to work).

OPTIONS FOR DEBUGGING

-D 0xN

--debug-mask=*0xN***]**

Specify a debug mask. For maximum debugging detail use **--debug-mask** = *0xffffffff*. To turn debugging messages off use **--debug-mask** = *0x0*. See debug.h for details. Note: that to compile without debugging also see debug.h.

--trace-summary

Keep track of information regarding system calls and print a summary. If tracing is enabled a list of each event being traced will be dumped to a trace output file.

--trace-summary-only

Keep track of information regarding system calls and print a summary. If this option is enabled individual events are not actually recorded but instead information to produce a summary of each event type is tracked. For example, minimum, maximum and average call times for each system call being traced.

-i

--interactive

Turns on interactive mode. The server stops between calls to process events and prompts the user for input. This permits the user to query some of the state of the server or to step through server execution. At the prompt type 'h' or '?' for a list of commands.

-P

--count-sigpipes

Install a signal handler that simply counts the number of SIGPIPE signals received.

OPTIONS FOR EXPERIMENTS RELATED TO BOUNDING PERFORMANCE

These options are designed to have the server stop processing a request at various stages of a requests life. It stops handling the request by closing it's end of the socket and not working on that request anymore. I've attempted to list these in the order of earliest time to drop the request to latest. The idea is that by running a series of experiments by using each of these options one might be able to get a sense of how much

additional stage costs.

-n

--no-accepts

--listen-only

This puts the server into a mode where it only ever listens to requests. It never makes any attempt to process requests. So it never even accepts any requests. The believe is that this might be used to provide insights into the amount of overhead incurred when kernel TCP SYN queues and application listen queues become full.

-A

--accepts-only

--close-after-accept

In this mode the server accepts and closes incoming connection requests as quickly as possible. At the time this was done *close()* would block until the call completed (despite the fd being in non blocking mode). So immediately after calling *accept()* the connection terminated by calling *close().*

--close-after-sock-init

The server accepts a connection, initializes the socket (setting socket options like socket buffer sizes, non blocking mode, etc) and then it closes the connection.

--close-after-read

The server accepts a connection, initializes the socket reads the request and then closes the connection.

--close-after-parse

The server accepts a connection, initializes the socket reads the request, parses the requests and then closes the connection.

OPTIONS FOR USE WITH THE SEND ENHANCED KERNEL

WARNING: these options are highly experimental and subject to continuous change.

These options specify how the **userver** interacts with new Linux Scalable Event Notification and Delivery (SEND) kernel mechanisms. See [Ostrowski-2000] for more details on SEND.

-K 0xN

--kdebug-mask=*0xN***]**

Specify a debugging mask which is passed to the *evtctl* system call to set a kernel debugging mask for the SEND related parts of the kernel.

--send-polls-for-accepts

The send loop polls for new connections.

-B

--block-for-send_events

When there are not more events to process, block and wait for one or more new events by calling *evtctl().*

-E

--send-events
Use send events. This version does not do notification.

-G

--use_getpid
If we aren't blocking to wait for events (by calling evtctl) we may wish to make a system call to give the kernel a chance to deliver events to the application (since they are delivered when returning from a syscall. In this case we use getpid() because it is likely fast.

-Y

--send-io-events
Use send events. This version does do notification.

-N

--send-loop
Use the send loop. Can potentially get and process send events even while using select. This bypasses the use of the select call.

-b N

--ecb-buf-entries=*N*
How many events can be stored in the application's event control buffer (ecb).

-Q N

--ecb-hi-water=*N*
Set the high water mark used for the event control buffer (ecb). One of the things we've tried is to turn event delivery off when the *--ecb-hi-water* mark is reached and then to turn delivery back on when the *--ecb-low-water* mark is reached.

-q N

--ecb-low-water=*N*
Set the low water mark used for the event control buffer (ecb). One of the things we've tried is to turn event delivery off when the *--ecb-hi-water* mark is reached and then to turn delivery back on when the *--ecb-low-water* mark is reached.

-U

--auto-accept
Turn on the SEND kernel's auto accept mechanisms. When new connection requests arrive the kernel automatically accepts them (auto accepts 'em) and notifies the application with an event containing the new connection's fd.

--use-accept-send
Use a special version of accept that returns event information related to the accepted connection/socket.

-j

--kernel-info

Experimental: was to tell the application that the kernel has an interface available for obtaining more information about the size and number of entries in some of the kernel queues (e.g., the TCP SYNQs and the application's listenq).

OUTPUT

This section describes portions of the output produced during the execution of the **userver.**

The first portion of output lists the program name and the command line parameters. The second portion of output lists the setting of all options. This is the most reliable way to find out the default settings for some options.

The next two sections identify operating system and per user or process limits.

The remainder of the output provides more information about default settings of various parameters and a large number of statistics about the execution of the server.

AUTHORS

userver was developed by Tim Brecht by starting with some micro web servers originally developed by Abhishek Chandra and David Mosberger [Chandra-2001]. David Pariag contributed the current caching engine. One of the hash functions is from Bob Jenkin's web site.

BUGS

Probably many.

NOTE: very few if any checks are performed to see if the combination of options being used makes sense. For example, the behaviour from using both the **--select-loop** and **--send-loop** options is undefined.

Always be sure to double-check results and don't fall prey to measuring client-performance instead of server performance! The tool **httperf** is very good at generating loads.

Many of the single letter options are difficult to relate to anything meaningful. As the number of options grew I eventually started to run out of alphabet and added long options.

The user-interface definitely could be improved. A simple configuration file might be more suitable than the many command-line options.

REFERENCES

[Brecht-2001] Tim Brecht and Michal Ostrowski, Exploring the Performance of Select-based Internet Servers, HP Labs Technical Report HPL-2001-314, November, 2001.

[Chandra-2001] A. Chandra and D. Mosberger, Scalability of Linux Event-Dispatch Mechanisms, Proceedings of the 2001 USENIX Annual Technical Conference, June 2001.

[Ostrowski-2000] Michal Ostrowski, A Mechanism for Scalable Event Notification and Delivery in Linux, M.Math Thesis, Department of Computer Science, University of Waterloo, November, 2000.