Integrating XMPP based communicator with large scale portal

Janusz Dziemidowicz

Erlang Factory Lite Kraków 2010

2nd of December 2010

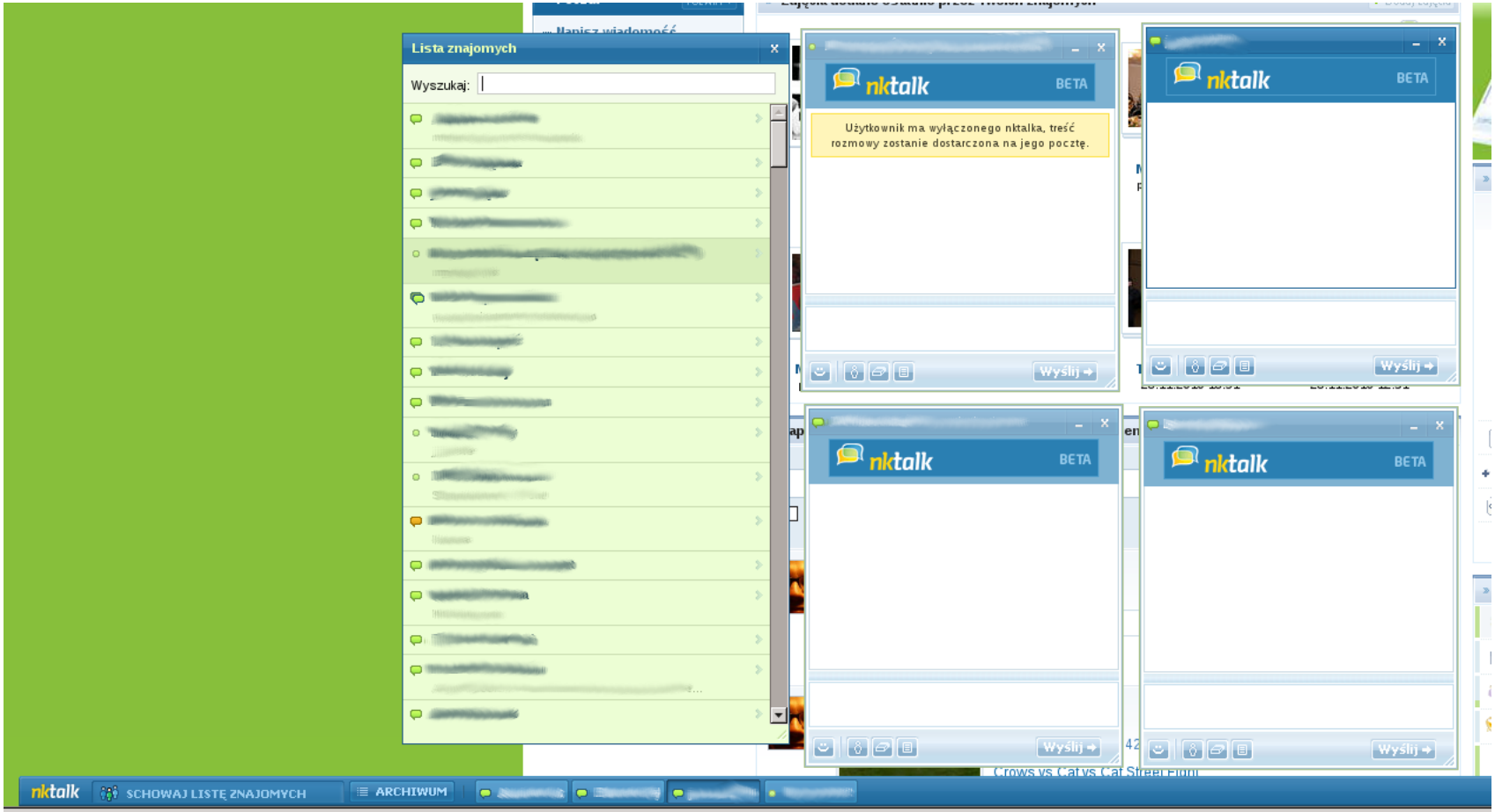
Table of contents

1. Why ejabberd?
   - Goal
   - Testing
2. Changes to ejabberd
   - Functional
   - Architectural and other
3. Problems we encountered
   - BOSH
   - pg2

Table of contents

1. Why ejabberd?
   - Goal
   - Testing
2. Changes to ejabberd
   - Functional
   - Architectural and other
3. Problems we encountered
   - BOSH
   - pg2

NkTalk - nk.pl IM

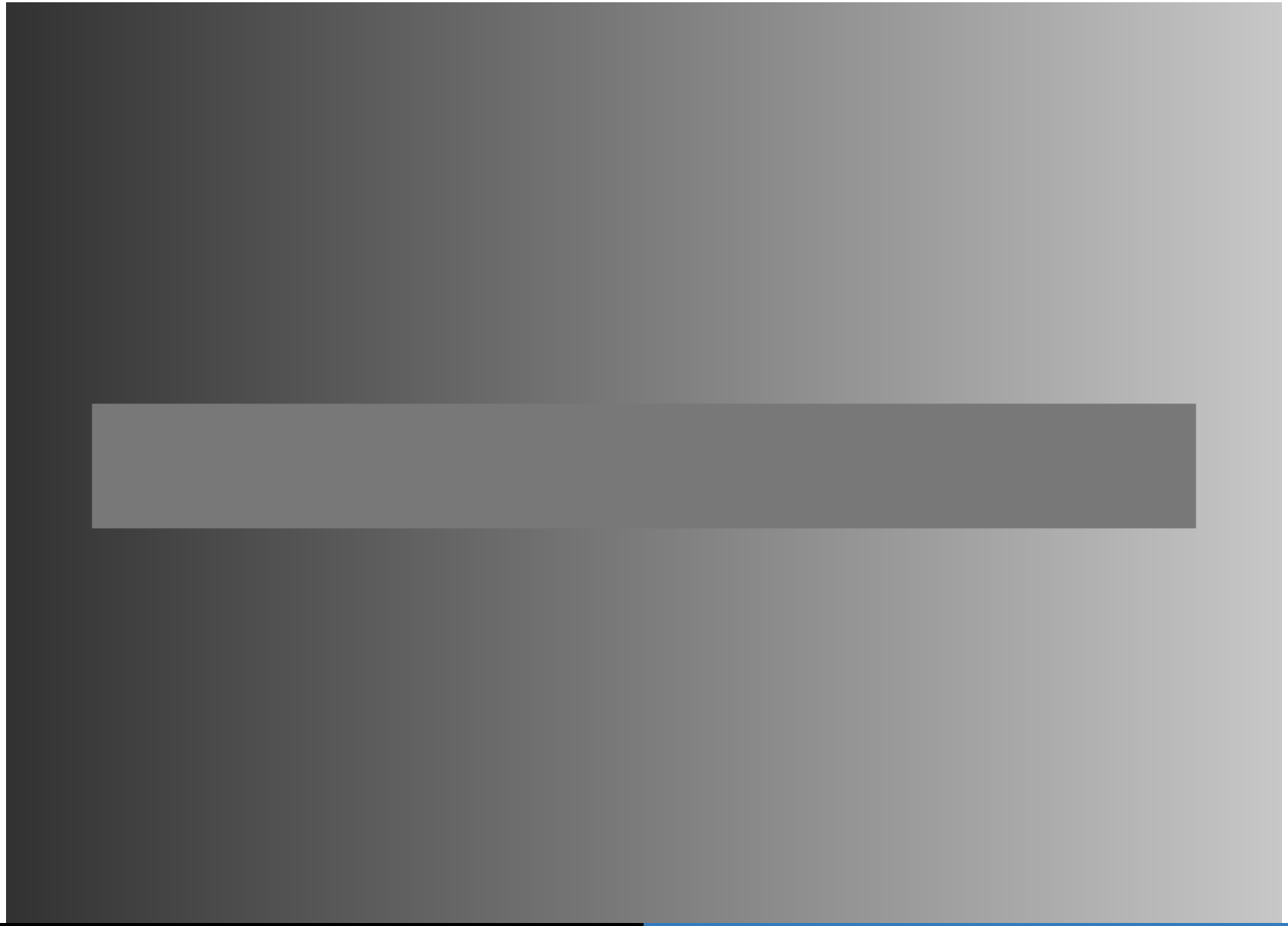


Possible solutions

Servers:

- Openfire,
- Tigase,
- ejabberd,
- write something from scratch.

First impressions



First impressions

Openfire:

- not suitable at all (sorry, no details).

Tigase:

- very nice performance,
- good contact with author, Artur Hefczyc.

ejabberd:

- completely unknown technology,
- SMP seemed to degrade performance,
- one node performed better than two nodes.

Writing from scratch:

- **lots** of work.

First impressions

Openfire:

- not suitable at all (sorry, no details).

Tigase:

- very nice performance,
- good contact with author, Artur Hefczyc.

ejabberd:

- completely unknown technology,
- SMP seemed to degrade performance,
- one node performed better than two nodes.

Writing from scratch:

- **lots** of work.

First impressions

Openfire:

- not suitable at all (sorry, no details).

Tigase:

- very nice performance,
- good contact with author, Artur Hefczyc.

ejabberd:

- completely unknown technology,
- SMP seemed to degrade performance,
- one node performed better than two nodes.

Writing from scratch:

- **lots** of work.

First impressions

Openfire:

- not suitable at all (sorry, no details).

Tigase:

- very nice performance,
- good contact with author, Artur Hefczyc.

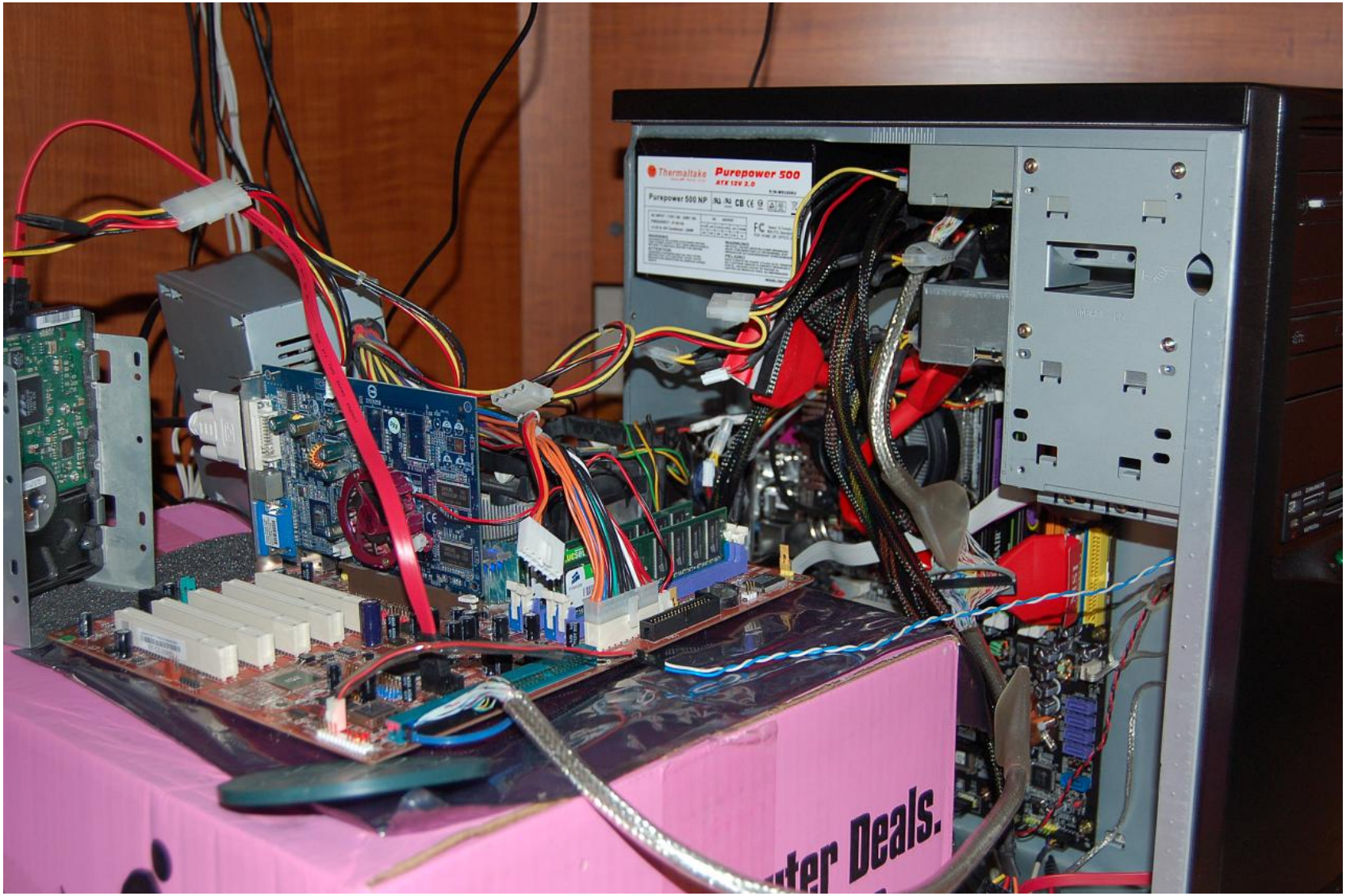
ejabberd:

- completely unknown technology,
- SMP seemed to degrade performance,
- one node performed better than two nodes.

Writing from scratch:

- **lots** of work.

Extensive testing



Extensive testing

- own testing framework written in Python (tsung was not enough),
- 32 servers with 8 cores for generating the load,
- 24 servers with 8 cores for running the XMPP servers,
- one million **online** users,
- 50,000 messages per second in peak.

Test results:

Tigase:

- beats ejabberd in performance when running on one machine,
- does not scale well, the more machines in cluster the worse.

ejabberd:

- http_bind in 2.0.5 was seriously broken,
- scales nicely (although not infinitely),
- requires some tuning like disabling unused modules, changing some Erlang parameters.

Both:

- support BOSH,
- same amount of work was needed to integrate with nk.pl.

Test results:

Tigase:

- beats ejabberd in performance when running on one machine,
- does not scale well, the more machines in cluster the worse.

ejabberd:

- http_bind in 2.0.5 was seriously broken,
- scales nicely (although not infinitely),
- requires some tuning like disabling unused modules, changing some Erlang parameters.

Both:

- support BOSH,
- same amount of work was needed to integrate with nk.pl.

Test results:

Tigase:

- beats ejabberd in performance when running on one machine,
- does not scale well, the more machines in cluster the worse.

ejabberd:

- http_bind in 2.0.5 was seriously broken,
- scales nicely (although not infinitely),
- requires some tuning like disabling unused modules, changing some Erlang parameters.

Both:

- support BOSH,
- same amount of work was needed to integrate with nk.pl.

Table of contents

1 Why ejabberd?
- Goal
- Testing

2 Changes to ejabberd
- Functional
- Architectural and other

3 Problems we encountered
- BOSH
- pg2


PREPARE
1
Relationships
2
Push it
3
CYA
4
Develop metrics
SET GOALS
Talk about business outcomes
Give people an opportunity to provide input
Tell people what they can do to help
BRING DONUTS OR BEER
BE COMPLETELY HONEST W/MY CO-WORKERS
Listen
TAKE IT TOO FAR AND THEN DIAL BACK
challenge assumptions
keep the momentum going
K.I.S.S.
speak up
CC the TRUE DECISION MAKER
CHANGE SUBJECT LINE IN EMAIL
ALWAYS INCLUDE CONTACT INFO IN EMAIL
Triple Check Spelling
TALK TO SHARON (LEADER)

Required ejabberd modules

- authentication,
- roster,
- privacy,
- offline,
- archive.

None of the original ejabberd modules were used.

Required ejabberd modules

- authentication,
- roster,
- privacy,
- offline,
- archive.

None of the original ejabberd modules were used.

MySQL support

We use MySQL a lot at nk.pl so we needed good support for it.

Things that we found lacking:

- horizontal partitioning,
- ability to change servers on the fly,
- good error handling.

All of the above had to be implemented.

MySQL support

We use MySQL a lot at nk.pl so we needed good support for it.

Things that we found lacking:

- horizontal partitioning,
- ability to change servers on the fly,
- good error handling.

All of the above had to be implemented.

MySQL support

We use MySQL a lot at nk.pl so we needed good support for it.

Things that we found lacking:

- horizontal partitioning,
- ability to change servers on the fly,
- good error handling.

All of the above had to be implemented.

Architectural changes

Architectural changes

Clustering:

- Mnesia stores only sessions and configuration (no “real” data),
- all Mnesia tables (including schema) are in RAM,
- automatic discovery and joining of new cluster nodes.

c2s:

- reworked presence handling to suit nk.pl needs,
- removed support for directed presences,
- reworked invisible presence,
- lower size of internal c2s state,
- lower size of session table entry.

Architectural changes

Clustering:

- Mnesia stores only sessions and configuration (no “real” data),
- all Mnesia tables (including schema) are in RAM,
- automatic discovery and joining of new cluster nodes.

c2s:

- reworked presence handling to suit nk.pl needs,
- removed support for directed presences,
- reworked invisible presence,
- lower size of internal c2s state,
- lower size of session table entry.

Other changes

- syslog support,
- extensive statistics via SNMP,
- various limits (number of presences, messages),
- custom listener to handle incoming events from portal.

Load balancing

- session is bound to particular server,
- binding is cookie based,
- we use haproxy as a HTTP load balancer.

In case of a node failure

Users bound to failed node are disconnected. After few seconds, web client tries to reconnect and haproxy directs all those users to a different node.

Load balancing

- session is bound to particular server,
- binding is cookie based,
- we use haproxy as a HTTP load balancer.

In case of a node failure

Users bound to failed node are disconnected. After few seconds, web client tries to reconnect and haproxy directs all those users to a different node.

Table of contents

1. Why ejabberd?
   - Goal
   - Testing
2. Changes to ejabberd
   - Functional
   - Architectural and other
3. Problems we encountered
   - BOSH
   - pg2

Problems with http_bind module

- retransmissions are not always handled properly,
- if user disconnects pending messages are not stored,
- crashes in some situations (usually on user disconnect),
- it was quite hard to fix those problems without reworking internals of the module.

Complete rewrite of http_bind

Advantages of our implementation:

- reliable handling of retransmissions in all cases (including out-of-order requests),
- stores pending messages in offline storage on user disconnect,
- properly working disconnect (instead of crash),
- adapts wait timeout dynamically,
- works as an ejabberd listener, bypassing ejabberd_http module,
- provides detailed statistics,
- conforms to latest versions of XEP-0124 and XEP-0206,
- ability to delay presences and group them in larger packets,
- hibernates on inactivity.

Complete rewrite of http_bind

Advantages of our implementation:

- reliable handling of retransmissions in all cases (including out-of-order requests),
- stores pending messages in offline storage on user disconnect,
- properly working disconnect (instead of crash),
- adapts wait timeout dynamically,
- works as an ejabberd listener, bypassing ejabberd_http module,
- provides detailed statistics,
- conforms to latest versions of XEP-0124 and XEP-0206,
- ability to delay presences and group them in larger packets,
- hibernates on inactivity.

Complete rewrite of http_bind

Disadvantages:

- doesn't support features we didn't need like session pausing, polling,
- not (yet) open source.

pg2 fail

pg2 fail

pg2

pg2 module is broken in R13B03 and R13B04. The more nodes in cluster the more duplicate processes appear in process group.

ejabberd

ejabberd uses pg2 internally for undocumented feature for splitting cluster into frontend and backend nodes.

fix

Just disable pg2 in ejabberd, it is not used under normal circumstances anyway.
Will be worked around in ejabberd 2.1.6 (EJAB-1349).

pg2 fail

pg2

pg2 module is broken in R13B03 and R13B04. The more nodes in cluster the more duplicate processes appear in process group.

ejabberd

ejabberd uses pg2 internally for undocumented feature for splitting cluster into frontend and backend nodes.

fix

Just disable pg2 in ejabberd, it is not used under normal circumstances anyway.
Will be worked around in ejabberd 2.1.6 (EJAB-1349).

pg2 fail

pg2

pg2 module is broken in R13B03 and R13B04. The more nodes in cluster the more duplicate processes appear in process group.

ejabberd

ejabberd uses pg2 internally for undocumented feature for splitting cluster into frontend and backend nodes.

fix

Just disable pg2 in ejabberd, it is not used under normal circumstances anyway.
Will be worked around in ejabberd 2.1.6 (EJAB-1349).

Thank you