

Connecting the data infrastructure with the DataFlow

Pere Urbon-Bayes

Software Architect

pere.urban@{gmail.com, acm.org}

Topics for Today

- Integration patterns for the *enterprise* startup.
- What is Apache NIFI.
- Examples
- NiFi on operation (best practises).

Integrate all the
things!

Enterprise integration is the task of **making separate applications work together** to produce an unified set of functionality.

The applications probably **run on multiple computers**, which may be geographically dispersed.

Some application **might need to be integrated even though they were not designed for integration** and can not be changed.

This issues, and others, are what makes application integration difficult.

Application coupling

Each **integration** faces different **needs** and criteria, we can group them as

Data formats and timeliness

Data or functionality

Communication

There is only a limited set of **integration options**



File transfer



Shared database

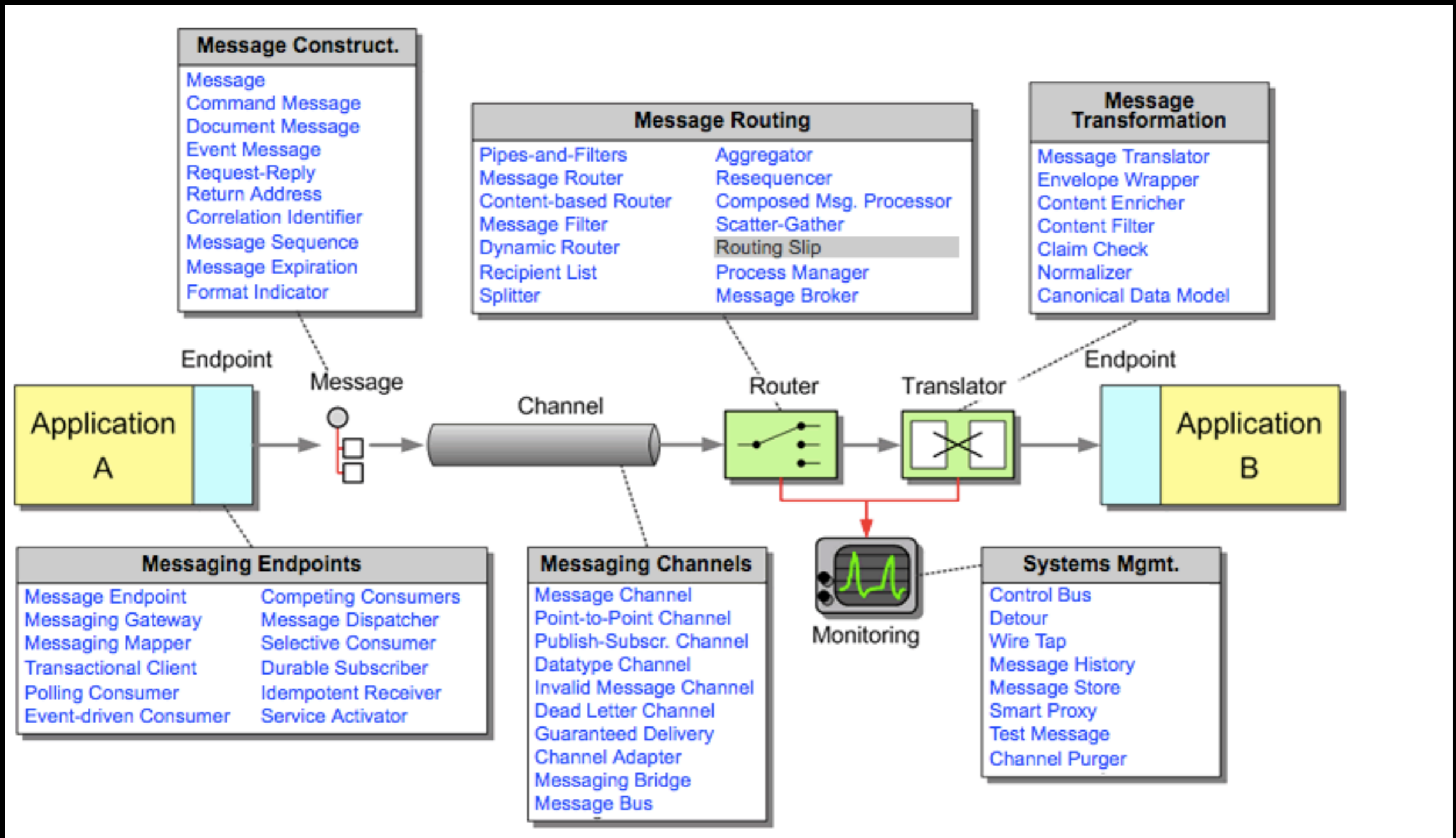


RPC invoke



Messaging

Enterprise Integration Patterns



What is Apache NiFi?

An *easy to use*, powerful, and reliable **system** to **process** and **distribute data**.

Web-based interface

Highly configurable

Data Provenance

Designed for extension

Secure

NiFi was **build to automate the flow of data between systems.**

But what is **Dataflow**?

an **automated and managed flow of information** between systems.

How Apache NiFi look like

The screenshot displays the Apache NiFi web interface. At the top, the browser address bar shows the URL: `localhost:8080/nifi/?processGroupId=root&componentIds=c0e732f1-015f-1000-54f1-ae9af7b4603`. The interface includes a navigation bar with various icons and a status bar showing the time as 19:13:27 CEST.

The main area shows a data flow diagram on a grid background. The flow starts with a **ConsumeKafka_0_10** processor, which connects to an **EvaluateJsonPath** processor. This processor has two outgoing flows: one labeled "Name success" and another labeled "Name matched". The "Name success" flow leads to another **EvaluateJsonPath** processor, which then connects to a **PutElasticsearchHttp** processor. The "Name matched" flow from the first **EvaluateJsonPath** processor leads to a **RouteOnAttribute** processor. This processor has three outgoing flows: "Name add", "Name update", and "Name delete". Each of these flows leads to a **PutElasticsearchHttp** processor.

On the left side, there are two panels: "Navigate" and "Operate". The "Operate" panel is currently displaying the configuration for the **EvaluateJsonPath** processor, showing its ID (`c0e732f1-015f-1000-54f1-ae9af7b4603`) and various control buttons like "DELETE".

Concepts behind Apache NiFi

A Flow file

Provenance Event

DETAILS ATTRIBUTES CONTENT

Attribute Values Show modified attributes only

filename
5468417820457

mime.type
application/gzip
No value set

path
./

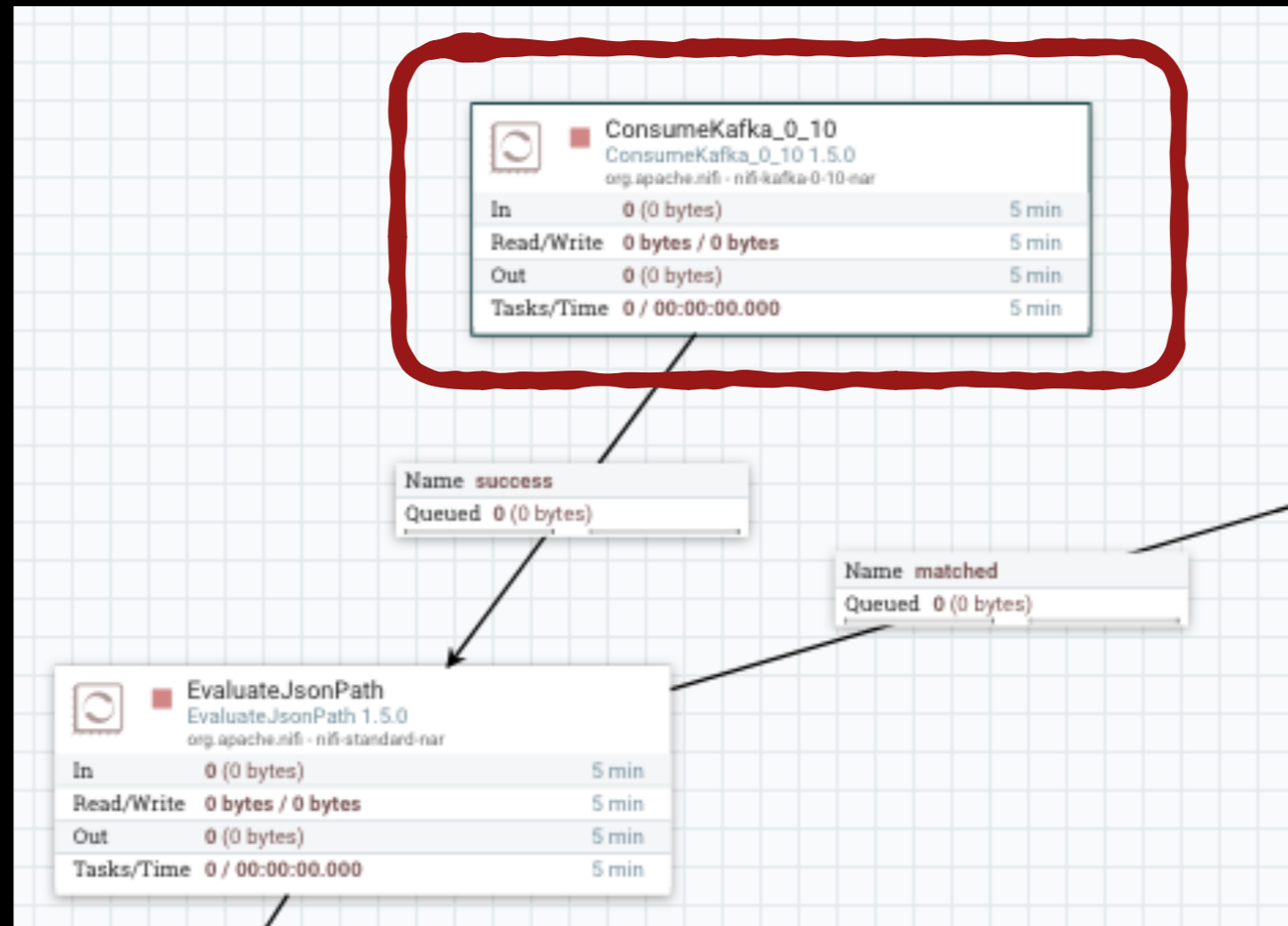
tcp.port
9696

tcp.sender
/127.0.0.1

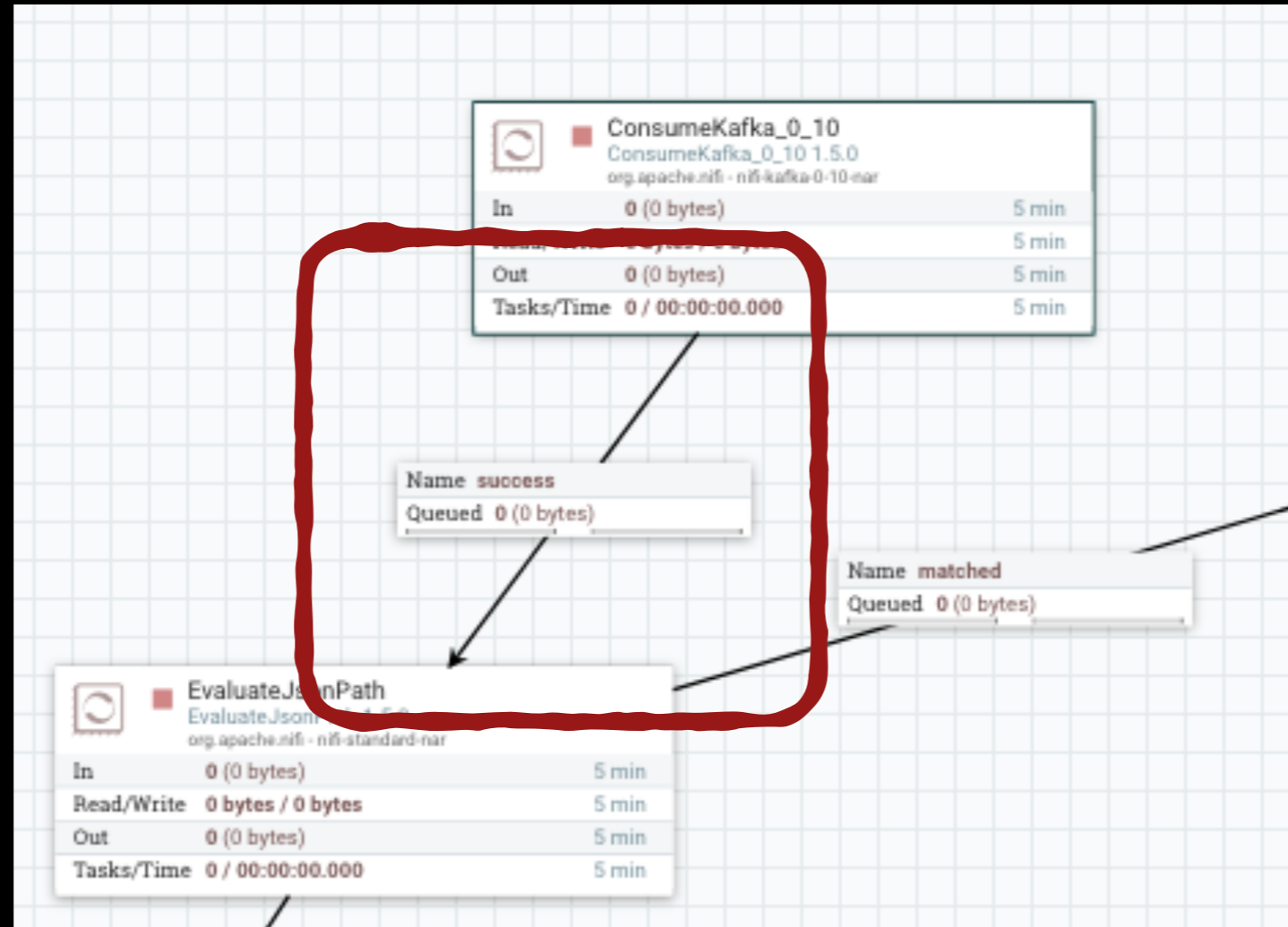
uuid
678fd60f-81ff-4f9c-9254-6974ddea2d13

OK


The Flow file Processor



A Connection



A Process Group

 **ListenHTTP**
ListenHTTP 1.5.0
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

To Receive Message

Name success

Queued 0 (0 bytes)

Demo Process Group


0 0 0 2 1 0

Queued	0 (0 bytes)	
In	0 (0 bytes) → 1	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	1 → 0 (0 bytes)	5 min

0 * 0 0 0 0 ? 0

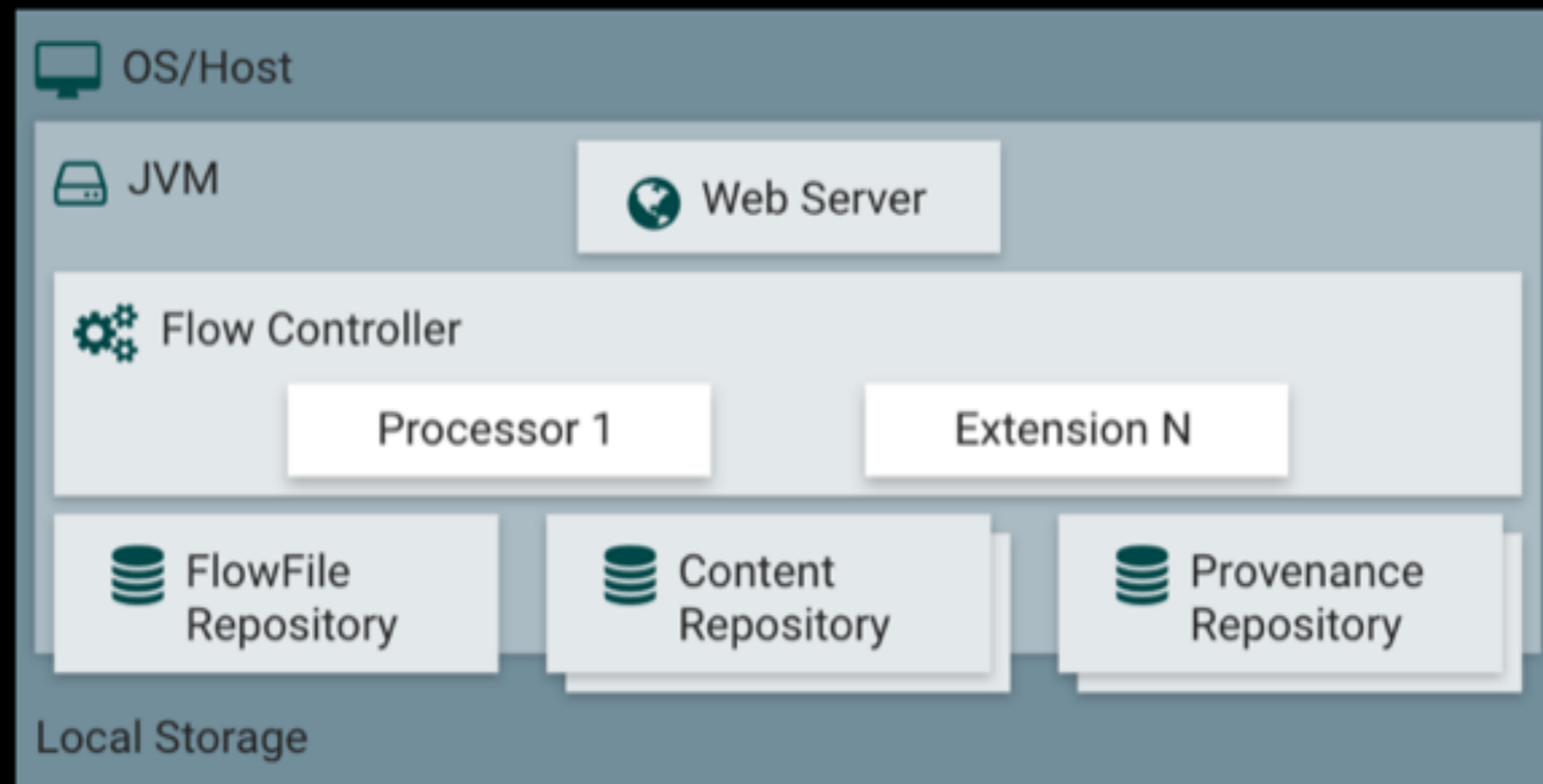
From sendMessage

Queued 0 (0 bytes)

 **SendEmail**
PutEmail 1.5.0
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Apache NiFi **Architecture**



Distributed using Apache Zookeeper

Let's take a closer
look...

Apache NiFi

Operations

```
/etc/security/limits.conf
```

```
hard nofile 50000
```

```
soft nofile 50000
```

Maximum file handles

```
/etc/security/limits.conf
```

```
hard nproc 10000
```

```
soft nproc 10000
```

```
/etc/security/limits.d/90-nproc.conf
```

Maximum forked Procs

```
sudo sysctl -w net.ipv4.ip_local_port_range="10000 65000"
```

Increase number of TCP sockets

```
sudo sysctl -w
```

```
net.ipv4.netfilter.ip_conntrack_tcp_timeout_time_wait="1"
```

Timeout sockets in TIMED_WAIT state

/etc/sysctl.conf

vm.swappiness = 0

/etc/fstab

/dev/sda7 /chroot ext2 defaults, **noatime** 1 2

Never SWAP

Thanks a lot!
Questions?
disagreements? threads?

Pere Urbon-Bayes
Data Wrangler
pere.urbon@{gmail.com, acm.org}