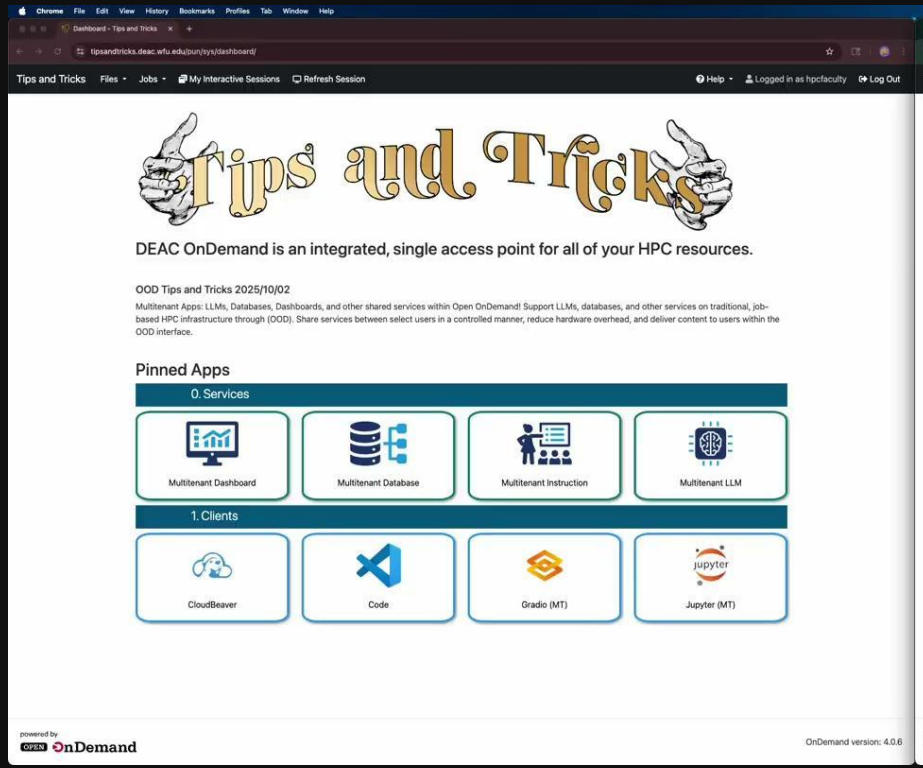
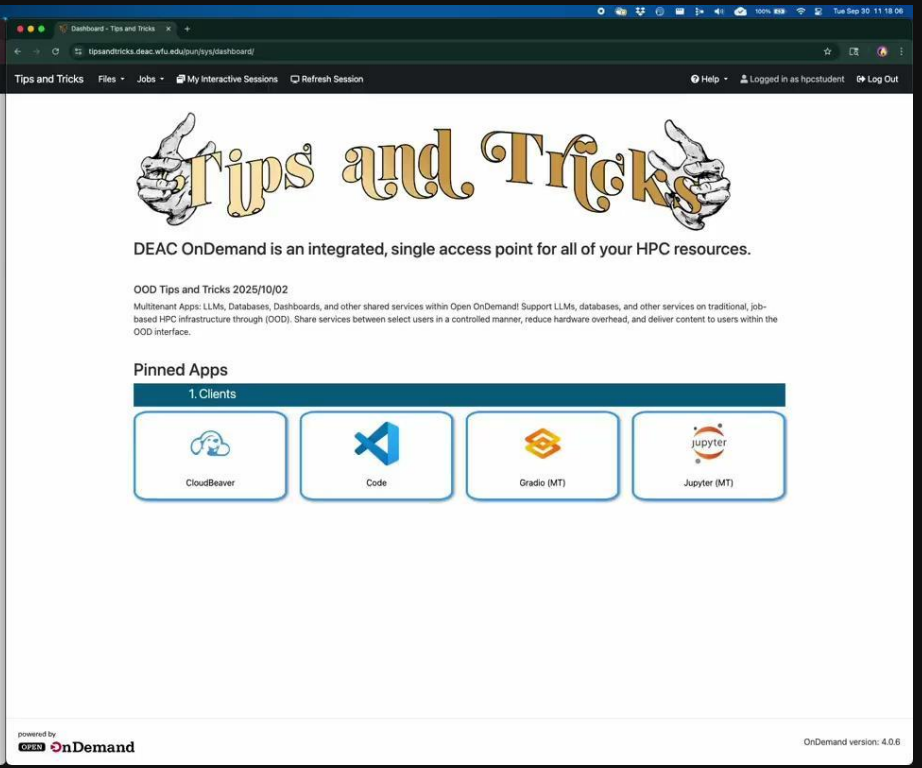
Multitenant Apps

Shared LLMs, Databases, and Other Services within OOD

WAKE FOREST UNIVERSITY

This presentation has no AI-generated content


Chrome File Edit View History Bookmarks Profiles Tab Window Help
Dashboard - Tips and Tricks
tipsandtricks.deac.wfu.edu/pun/sys/dashboard/
Tips and Tricks
Files
Jobs
My Interactive Sessions
Refresh Session
Help
Logged in as hpcfaculty
Log Out
Tips and Tricks
DEAC OnDemand is an integrated, single access point for all of your HPC resources.
OOD Tips and Tricks 2025/10/02
Multitenant Apps: LLMs, Databases, Dashboards, and other shared services within Open OnDemand! Support LLMs, databases, and other services on traditional, job-based HPC infrastructure through (OOD). Share services between select users in a controlled manner, reduce hardware overhead, and deliver content to users within the OOD interface.
Pinned Apps
0. Services
Multitenant Dashboard
Multitenant Database
Multitenant Instruction
Multitenant LLM
1. Clients
CloudBeaver
Code
Gradio (MT)
Jupyter (MT)
powered by
OPEN OnDemand
OnDemand version: 4.0.6


Dashboard - Tips and Tricks
tipsandtricks.deac.wfu.edu/pun/sys/dashboard/
Tips and Tricks
Files
Jobs
My Interactive Sessions
Refresh Session
Help
Logged in as hpcstudent
Log Out
Tips and Tricks
DEAC OnDemand is an integrated, single access point for all of your HPC resources.
OOD Tips and Tricks 2025/10/02
Multitenant Apps: LLMs, Databases, Dashboards, and other shared services within Open OnDemand! Support LLMs, databases, and other services on traditional, job-based HPC infrastructure through (OOD). Share services between select users in a controlled manner, reduce hardware overhead, and deliver content to users within the OOD interface.
Pinned Apps
1. Clients
CloudBeaver
Code
Gradio (MT)
Jupyter (MT)
powered by
OPEN OnDemand
OnDemand version: 4.0.6

Chrome File Edit View History Bookmarks Profiles Tab Window Help

My Interactive Sessions - Tic

tipsandtricks.deac.wfu.edu/pun/sys/dashboard/batch_connect/sessions

Tips and Tricks | Files | Jobs | My Interactive Sessions | Refresh Session | Help | Logged in as hpcfaculty | Log Out

Session was successfully created.

Home / My Interactive Sessions

Multitenant Apps

Apps

- Multitenant Dashboard
- Multitenant Database
- Multitenant Instruction
- Multitenant LLM

Clients

- Jupyter (MT)
- Code
- CloudBeaver
- Gradio (MT)

Multitenant LLM (5689064) 1 node | 64 cores | Running

Cancel

Host: gpu-a100-03.deac.wfu.edu

Created at: 2025-09-30 11:19:03 EDT

Time Remaining: 51 minutes

Session ID: a68041d9-cf2e-478b-81a7-a0772225c344

Problems with this session? Submit support ticket

Ollama API URL:

```
http://gpu-a100-03.deac.wfu.edu:52587
```

Launch Code Server | Launch Jupyter

powered by OPEN OnDemand

OnDemand version: 4.0.6

My Interactive Sessions - Tic

tipsandtricks.deac.wfu.edu/pun/sys/dashboard/batch_connect/sessions

Tue Sep 30 11:27:27

Tips and Tricks | Files | Jobs | My Interactive Sessions | Refresh Session | Help | Logged in as hpcstudent | Log Out

Home / My Interactive Sessions

Multitenant Apps

Clients

- Jupyter (MT)
- Code
- CloudBeaver
- Gradio (MT)

Ollama_API from hpcfaculty (5689064) 1 node | 64 cores | Running

Cancel

Host: gpu-a100-03.deac.wfu.edu

Created at: 2025-09-30 11:19:32 EDT

Time Remaining: 51 minutes

Session ID: a68041d9-cf2e-478b-81a7-a0772225c344

Problems with this session? Submit support ticket

Ollama API URL:

```
http://gpu-a100-03.deac.wfu.edu:52587
```

The Ollama API URL shown above has been automatically set in the `$OLLAMA_HOST` environment variable.

For Jupyter notebooks: Select the Ollama API from the list in the form. Once in the Notebook, list the models available like this:

```
import ollama
for m in ollama.list().models:
  print(m.model)
```

Launch Jupyter | Launch Gradio

powered by OPEN OnDemand

OnDemand version: 4.0.6

Goals/Use Cases

We want to:

1. Offer an LLM software stack that select users can share and access from within the cluster
2. Offer a database software stack that select users can share and access from within the cluster
3. Enable PIs to share dashboards and other web services within their research groups or departments
4. ??? (Bonus)

but using traditional job-based HPC infrastructure!

Inspiration (Local)

- **LLM**: Spanish faculty that wants local LLM with published material
- **LLM**: Software engineering class that uses gpt-oss LLM for agentic coding with Cline
- **Databases**: New MSBA (Business Analytics) program
- **Databases**: New Athletics Data Analytics vertical

Inspiration (Community)

- Conversation with Travis Ravert at GOOD25 about LLM backends
- April's Tips and Tricks presentation by Ron Rahaman from Georgia Tech's PACE
- PEARC24 paper on Stable Diffusion in the Classroom

Chrome File Edit View History Bookmarks Profiles Tab Window Help

My Interactive Sessions - Tip

tipsandtricks.deac.wfu.edu/pun/sys/dashboard/batch_connect/sessions

Tips and Tricks | Files | Jobs | My Interactive Sessions | Refresh Session | Help | Logged in as hpcfaculty | Log Out

Home / My Interactive Sessions

Multitenant Apps

Apps

- Multitenant Dashboard
- Multitenant Database
- Multitenant Instruction
- Multitenant LLM

Clients

- Jupyter (MT)
- Code
- CloudBeaver
- Gradio (MT)

Multitenant LLM (5689090) 1 node | 64 cores | Running

Host: gpu-a100-03.deac.wfu.edu

Cancel

Created at: 2025-09-30 11:34:15 EDT

Time Remaining: 55 minutes

Session ID: 0b91935a-95e7-4e52-8037-a5ba3a451827

Problems with this session? Submit support ticket

Ollama API URL:

```
http://gpu-a100-03.deac.wfu.edu:34833
```

Launch Code Server | Launch Jupyter

powered by OPEN OnDemand

OnDemand version: 4.0.6

Multitenant LLM + Jupyter

Chrome File Edit View History Bookmarks Profiles Tab Window Help 100% Tue Sep 30 11 41 30

My Interactive Sessions - Tip

tipsandtricks.deac.wfu.edu/pun/sys/dashboard/batch_connect/sessions

Tips and Tricks | Files | Jobs | My Interactive Sessions | Refresh Session | Help | Logged in as hpcstudent | Log Out

Home / My Interactive Sessions

Multitenant Apps

Clients

- Jupyter (MT)
- Code
- CloudBeaver
- Gradio (MT)

Ollama_API from hpcfaculty (5689090) 1 node | 64 cores | Running

Host: gpu-a100-03.deac.wfu.edu

Cancel

Created at: 2025-09-30 11:34:39 EDT

Time Remaining: 52 minutes

Session ID: 0b91935a-95e7-4e52-8037-a5ba3a451827

Problems with this session? Submit support ticket

Ollama API URL:

```
http://gpu-a100-03.deac.wfu.edu:34833
```

The Ollama API URL shown above has been automatically set in the `$OLLAMA_HOST` environment variable.

For Jupyter notebooks: Select the Ollama API from the list in the form. Once in the Notebook, list the models available like this:

```
import ollama
for m in ollama.list().models:
  print(m.model)
```

Launch Jupyter | Launch Gradio

Multitenant LLM + Gradio

Multitenant Apps

Shared LLMs, Databases, and Other Services within OOD

WAKE FOREST UNIVERSITY

This presentation has no AI-generated content

Ollama_Code from hpcfaculty (5689095)

1 node | 64 cores | Running

Host: gpu-a100-03.deac.wfu.edu

Cancel

Created at: 2025-09-30 11:44:34 EDT

Time Remaining: 57 minutes

Session ID: 9b669527-1941-40d3-97dc-d9b7383d7b85

Problems with this session? Submit support ticket

Make sure you have the Cline extension installed within Code Server. Use the following parameters to configure Cline:

Use your own API key
API Provider: `Ollama`
Use custom base URL: ✓
Base URL: `http://gpu-a100-03.deac.wfu.edu:37041`
Model: `gpt-oss:120b`

Launch Code Server

Multitenant LLM for AI Coding

DB_Conn from hpcfaculty (5689101)

1 node | 1 core | Running

Host: cpu-amd-24.deac.wfu.edu

Cancel

Created at: 2025-09-30 12:02:28 EDT

Time Remaining: 59 minutes

Session ID: 263a9e8a-e0db-4bb7-90aa-96a3fa411844

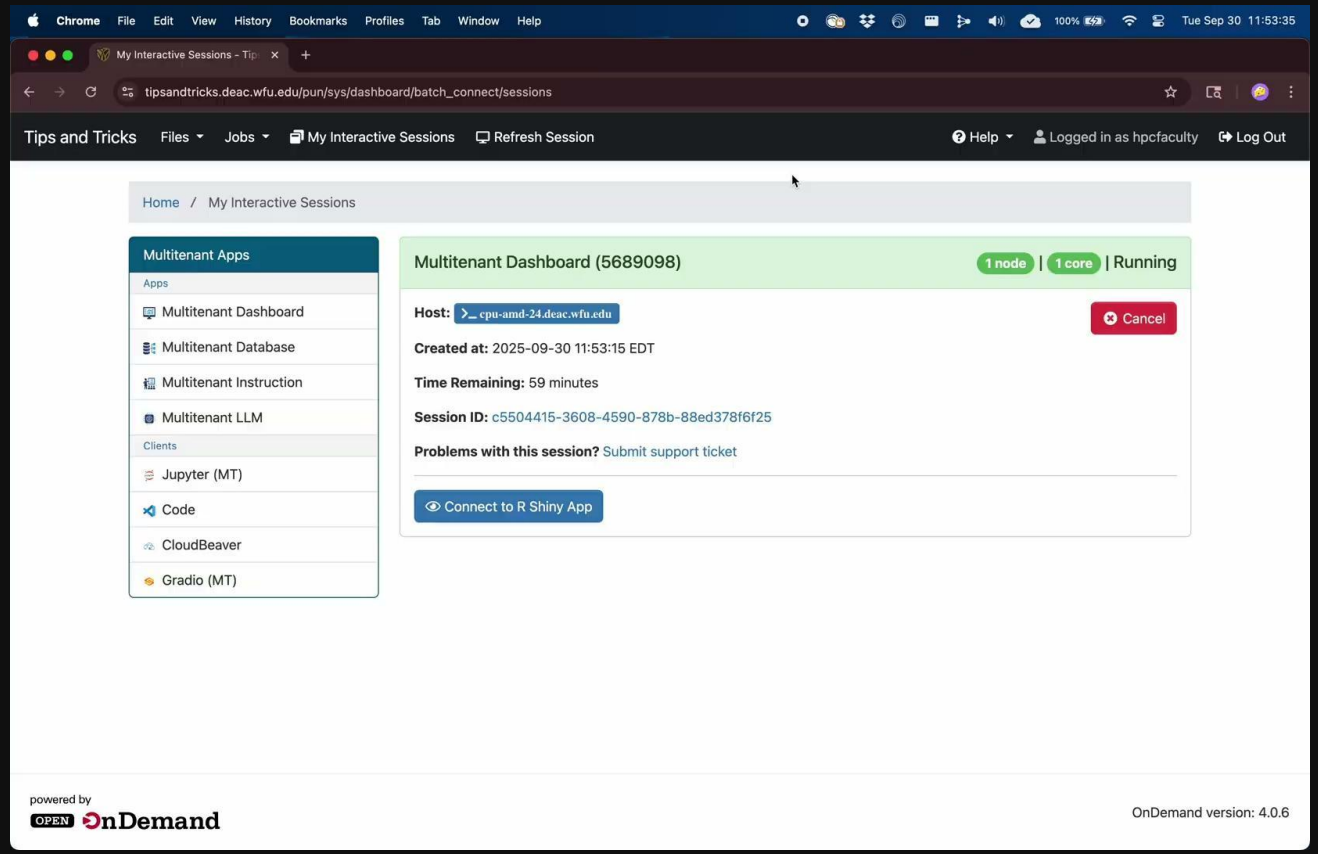
Problems with this session? Submit support ticket

Database Connection Details

Host: `cpu-amd-24.deac.wfu.edu`
Port: `44749`
Database: `pagila`
Username: `deac`
Password: `godeacs`

Launch CloudBeaver | Launch DBeaver

Multitenant Database



Multitenant Dashboard (Web Server)

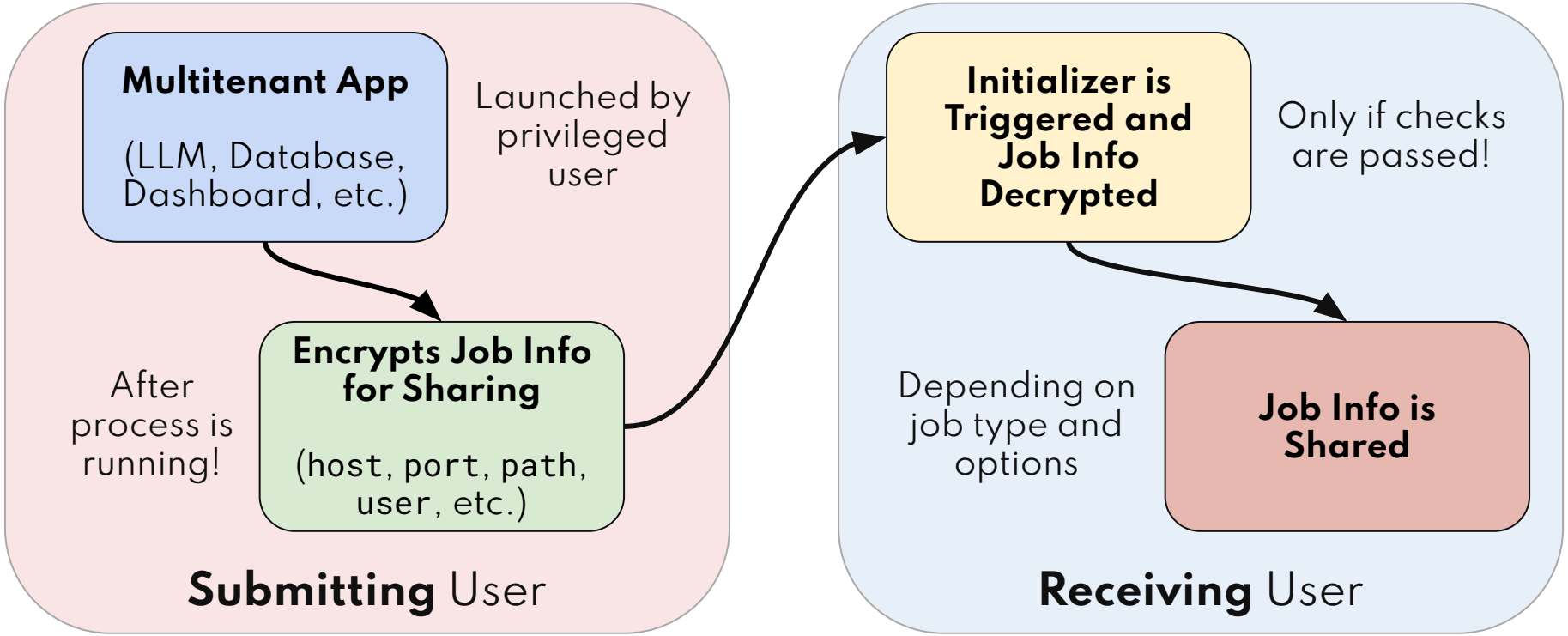
What these are NOT

- Not for enterprise applications
- Not for high security applications
- Not for sharing Jupyter Notebooks, RStudio or other “log-in” or user-facing applications
- Not for sharing interactive VNC connections (maybe view-only)

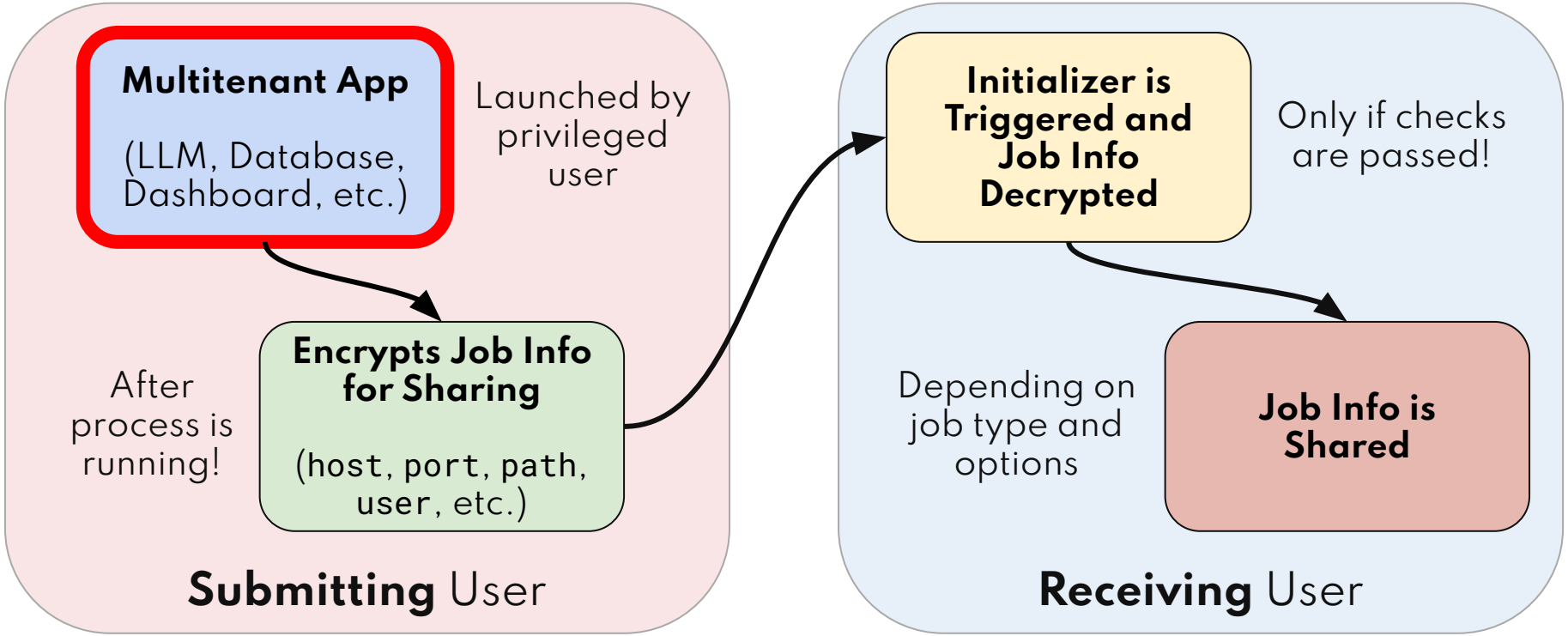
Requirements

- Slurm Scheduler (post 2014)
- Compute nodes need to talk to each other
- `openssl`
- `gzip`

Overview


Multitenant App
(LLM, Database, Dashboard, etc.)
Launched by privileged user
Encrypts Job Info for Sharing
(host, port, path, user, etc.)
After process is running!
Submitting User
Initializer is Triggered and Job Info Decrypted
Only if checks are passed!
Job Info is Shared
Depending on job type and options
Receiving User

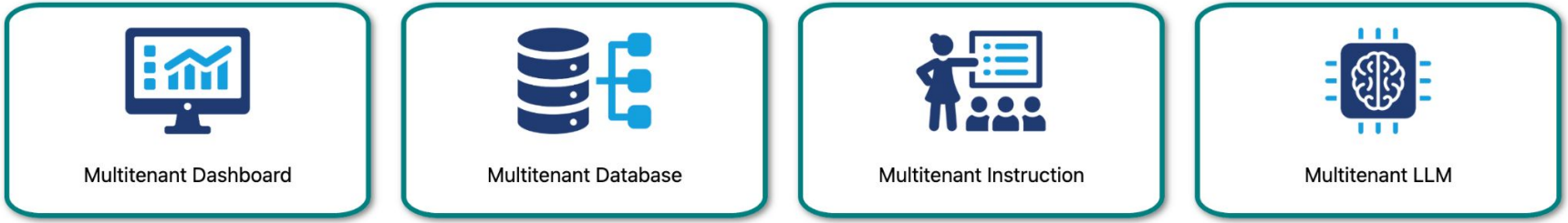
Overview


Multitenant App
(LLM, Database, Dashboard, etc.)
Launched by privileged user
Encrypts Job Info for Sharing
(host, port, path, user, etc.)
After process is running!
Submitting User
Initializer is Triggered and Job Info Decrypted
Only if checks are passed!
Job Info is Shared
Depending on job type and options
Receiving User

The Multitenant App: Part 0

Multitenant App

(LLM, Database, Dashboard, etc.)



The Multitenant App: Part 0

Multitenant App

(LLM, Database, Dashboard, etc.)

```
apps
└── multitenant-llm
    ├── manifest.yml
    ├── view.html.erb
    ├── form.yml.erb
    ├── submit.yml.erb
    └── template
        ├── after.sh.erb
        ├── before.sh.erb
        └── script.sh.erb
```

The Multitenant App: Part 0

Multitenant App

(LLM, Database, Dashboard, etc.)

```
export ollama_url="http://${host}:${port}"
export ollama_models="/data/ollama"
export ollama_model="gpt-oss:120b”

export OLLAMA_MODELS=$ollama_models
export OLLAMA_HOST=$ollama_url
export OLLAMA_ORIGINS="*"
export OLLAMA_SCHED_SPREAD=1
export OLLAMA_TMPDIR=/scratch/${SLURM_JOBID}
export FORWARDED_ALLOW_IPS="127.0.0.1"

# launch Ollama
ollama serve
```

The Multitenant App: Part 0

Multitenant App

(LLM, Database, Dashboard, etc.)

Multitenant App Options

No Multitenancy

Do you want your job to be shared with select users?

Runs your App with no sharing.
Business as usual!

The Multitenant App: Part 0

Multitenant App

(LLM, Database, Dashboard, etc.)

Multitenant App Options

My User Only

Do you want your job to be shared with select users?

Passes the shared info into other Interactive Apps.

Your user only!

```
submitting user == receiving user
```

The Multitenant App: Part 0

Multitenant App

(LLM, Database, Dashboard, etc.)

Multitenant App Options

Other Users

Do you want your job to be shared with select users?

Tenant Group

hpcGrp

Select the group of users that should connect to your app.

Delivery Method

Form View Only

What type of view do you want your users to see?

Ready to share with other users

```
submitting user != receiving user
```

The Multitenant App: Part 0

Multitenant App

(LLM, Database, Dashboard, etc.)

Multitenant App Options

Other Users

Do you want your job to be shared with select users?

Tenant Group

hpcGrp

Select the group of users that should connect to your app.

Delivery Method

Form View Only

What type of view do you want your users to see?

Choose the POSIX group that you want to share with.

The Multitenant App: Part 0

Multitenant App

(LLM, Database, Dashboard, etc.)

Multitenant App Options

Other Users

Do you want your job to be shared with select users?

Tenant Group

hpcGrp

Select the group of users that should connect to your app.

Delivery Method

Form View Only

What type of view do you want your users to see?

Choose how you want to deliver the info:

- **Form view**: available directly in other apps

The Multitenant App: Part 0

Multitenant App

(LLM, Database, Dashboard, etc.)

Multitenant App Options

Other Users

Do you want your job to be shared with select users?

Tenant Group

hpcGrp

Select the group of users that should connect to your app.

Delivery Method

Card + Form View

What type of view do you want your users to see?

Choose how you want to deliver the info:

- **Form view**: available directly in other apps
- **Card view**: content delivered to dashboard

The Multitenant App: Part 0

Multitenant App

(LLM, Database, Dashboard, etc.)

Delivery Method

Card + Form View

What type of view do you want your users to see?

Card Preset

Default Ollama API (Jupyter/Python)

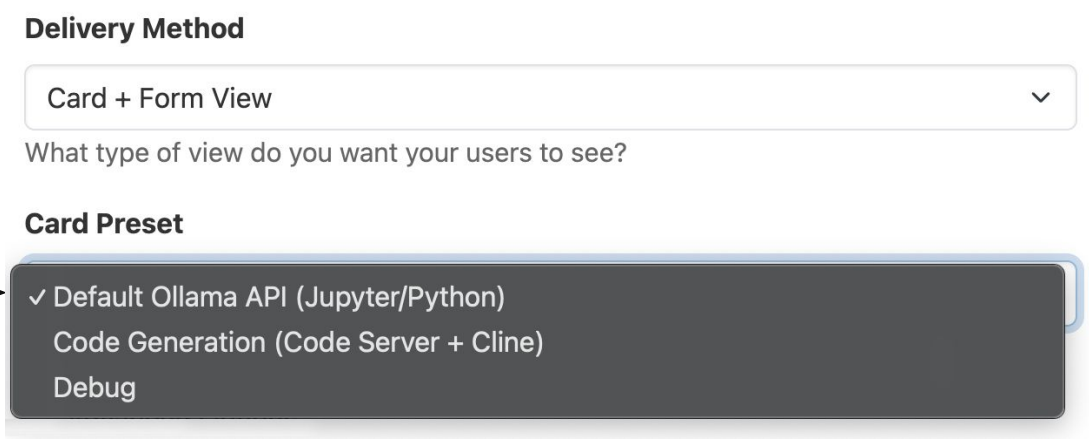
The style of card you want delivered to tenants.

Choose the preset for displaying the info

The Multitenant App: Part 0

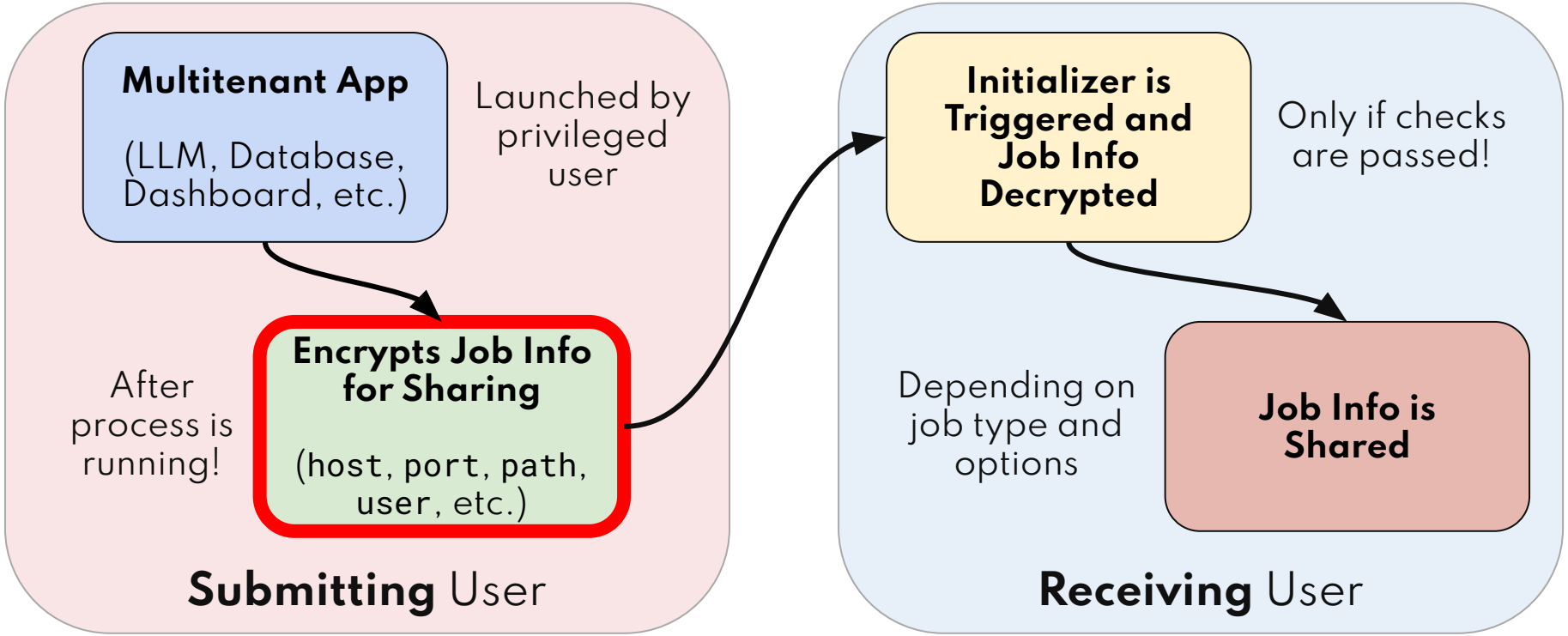
Multitenant App

(LLM, Database, Dashboard, etc.)



Choose the preset for displaying the info

Overview


Multitenant App
(LLM, Database, Dashboard, etc.)
Launched by privileged user
Encrypts Job Info for Sharing
(host, port, path, user, etc.)
After process is running!
Submitting User
Initializer is Triggered and Job Info Decrypted
Only if checks are passed!
Job Info is Shared
Depending on job type and options
Receiving User

The Multitenant App: Part 1

```
apps
└── multitenant-llm
    ├── manifest.yml
    ├── view.html.erb
    ├── form.yml.erb
    ├── submit.yml.erb
    └── template
        ├── after.sh.erb
        ├── before.sh.erb
        └── script.sh.erb
```

Encrypts Job Info for Sharing

(host, port, path, user, etc.)

The Multitenant App: Part 1

```
$ cat submit.yml.erb

batch_connect:
  template: basic
  conn_params:
    - ollama_url
    - ollama_models
    - ollama_model
```

Encrypts Job Info for Sharing

(host, port, path, user, etc.)

The Multitenant App: Part 1

Encrypts Job Info for Sharing

(host, port, path, user, etc.)

```
$ cat after.sh.erb

mt_connection=$(cat <<EOF
{
  'host':              '${host}',
  'ollama_url':        '${ollama_url}',
  'ollama_models':     '${ollama_models}',
  'ollama_model':      '${ollama_model}'
}
EOF
)
```

The Multitenant App: Part 1

Encrypts Job Info for Sharing

(host, port, path, user, etc.)

```
# continued from previous slide

mt_accounting=$(cat << EOF
{
  'mtu':  'uid001,uid002,uid003,...,uid150',
  'mti':  '${sessionid}',
  'mta':  '<%= context.mt_appname %>',
  'mtm':  '<%= context.mt_method %>',
  'mtd':  '<%= context.mt_delivery %>'
}
EOF
)
```

The Multitenant App: Part 1

Encrypts Job Info for Sharing

(host, port, path, user, etc.)

```
# continued from previous slide

mt_message="${mt_accounting}${mt_connection}"

ngroup="<%= context.auto_groups %>"

nmessage=$(encrypt "$mt_message" "$key" "$iv")

nfinal="sys/dashboard|${ngroup}|${nmessage}"
```

???

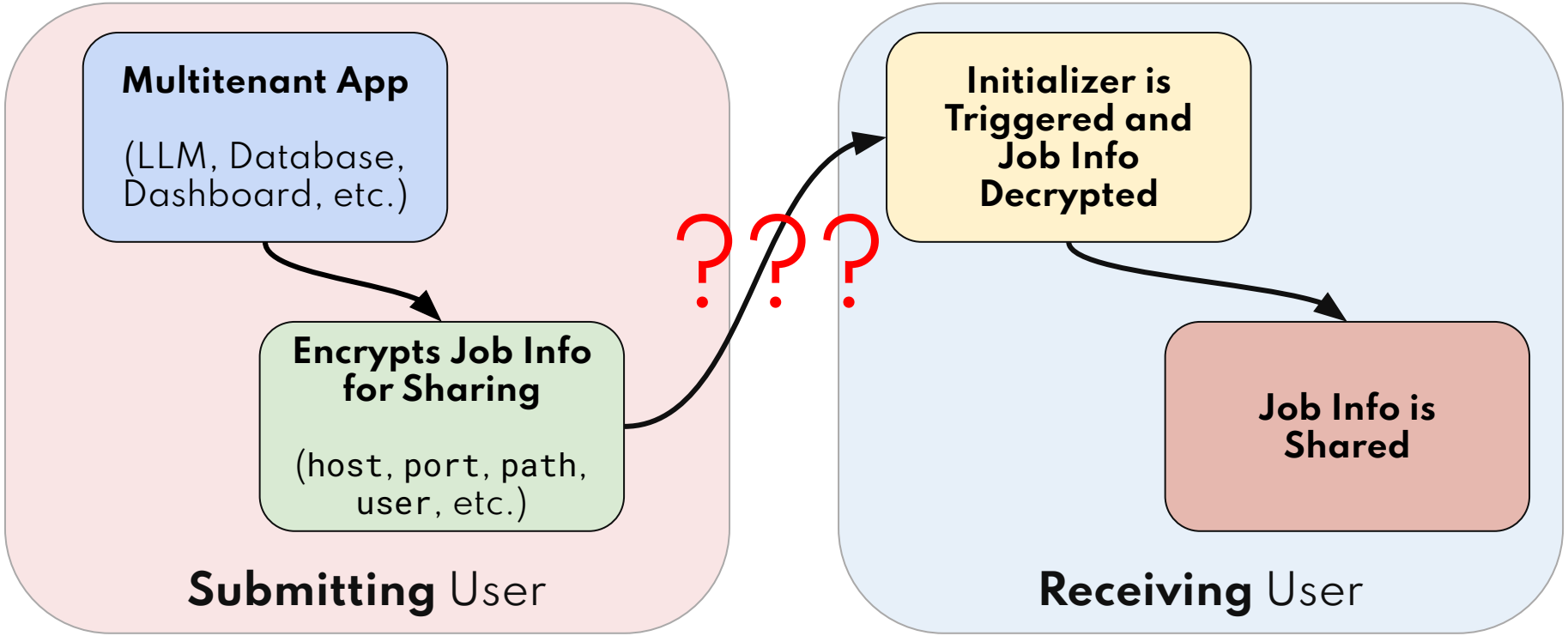
The Multitenant App: Part 1

```
scontrol update \
  JobId=${SLURM_JOB_ID} \
  jobname="${nfinal}"
```

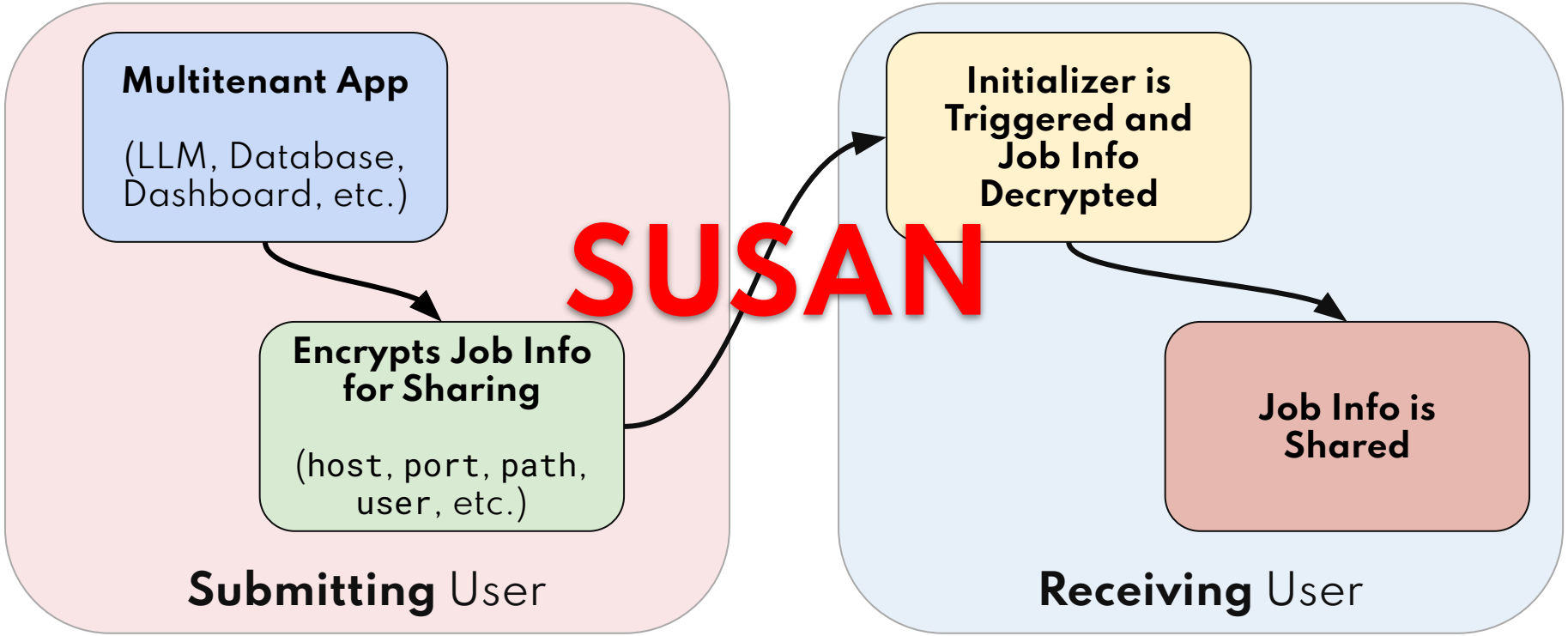
Encrypts Job Info for Sharing

(host, port, path, user, etc.)

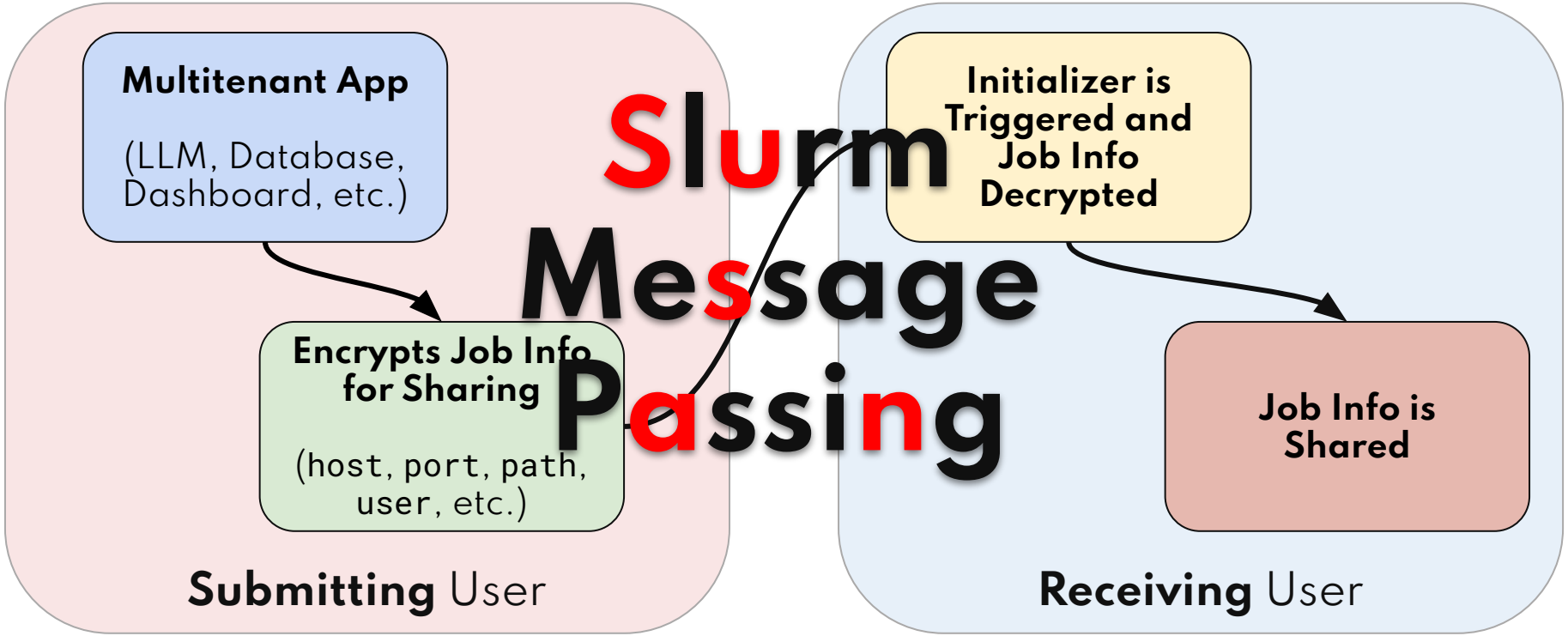
Overview


Multitenant App
(LLM, Database, Dashboard, etc.)
Encrypts Job Info for Sharing
(host, port, path, user, etc.)
Submitting User
???
Initializer is Triggered and Job Info Decrypted
Job Info is Shared
Receiving User

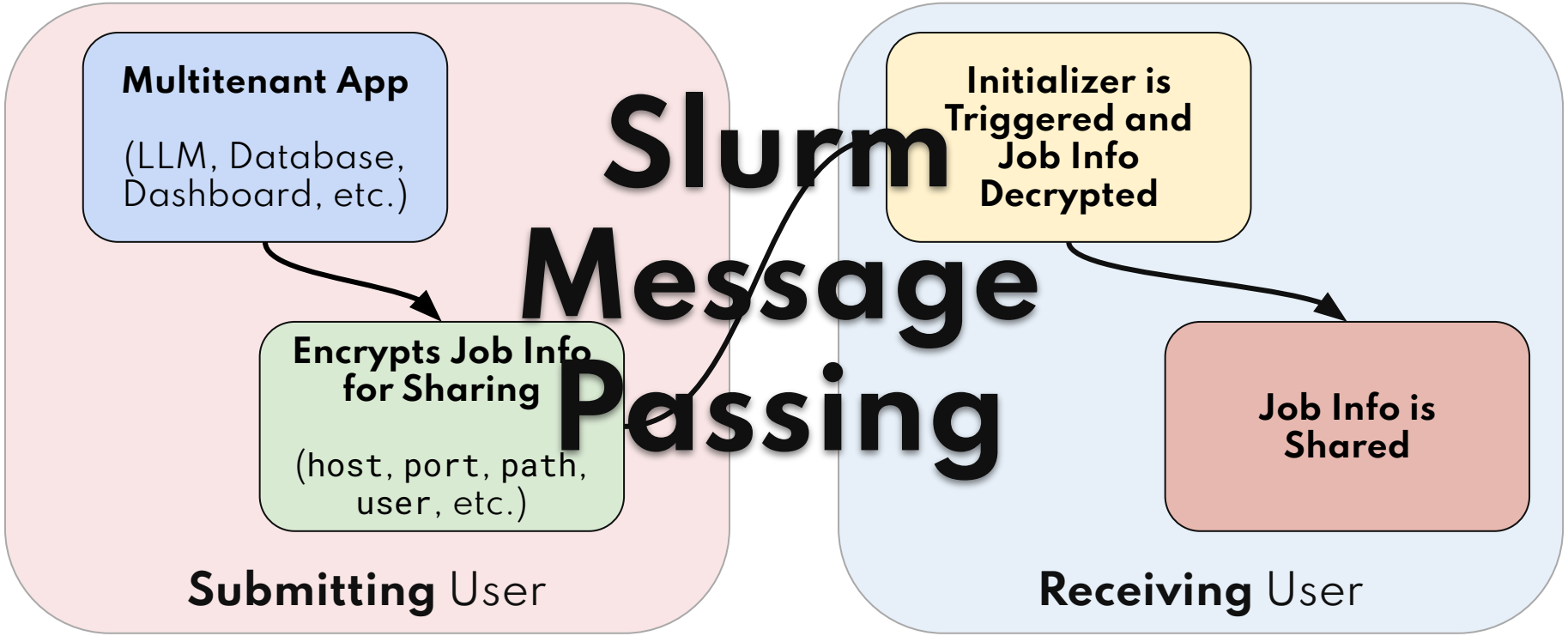
Overview


Multitenant App
(LLM, Database, Dashboard, etc.)
Encrypts Job Info for Sharing
(host, port, path, user, etc.)
Submitting User
SUSAN
Initializer is Triggered and Job Info Decrypted
Job Info is Shared
Receiving User

Overview


Multitenant App
(LLM, Database, Dashboard, etc.)
Encrypts Job Info for Sharing
(host, port, path, user, etc.)
Submitting User
Slurm Message Passing
Initializer is Triggered and Job Info Decrypted
Job Info is Shared
Receiving User

Overview


Multitenant App
(LLM, Database, Dashboard, etc.)
Encrypts Job Info for Sharing
(host, port, path, user, etc.)
Submitting User
Slurm Message Passing
Initializer is Triggered and Job Info Decrypted
Job Info is Shared
Receiving User

Why Slurm Message Passing?

- Slurm is already present everywhere on the cluster
- Slurm has built-in info about every job
  - jobid, host, submitter, etc.
- Slurm queue requires no special permissions to view
- Slurm and OOD interact directly with each other
- **Job name:**
  - Ubiquitous across all Slurm jobs and installations!
  - Job owner can modify after submission!

Why Slurm Message Passing?

Unique Slurm feature: **WCKeys**

> "A WCKey is an orthogonal way to do accounting against possibly unrelated accounts. This can be useful where users from different accounts are all working on the same project."

Minimal modification to your Slurm config!*

Why Slurm Message Passing?

Unique Slurm feature: **WCKeys**

```
submit.yml.erb:
    script:
      native:
        ...
         - "--wckey=multitenant"
```

Why Slurm Message Passing?

vs. writing a file:

- Availability on filesystem:
  - Can we guarantee that the file will be there?
- File permissions:
  - Can we guarantee the user can read it?
- Cleanup:
  - Can we avoid cruft and data leakage?

Why Slurm Message Passing?

Limitation:

1024 character limit!

What’s in a name?

sys/dashboard|hpcGrp|FTYql6YwioiGTCTaEKHY686l6ZtQLnv/m/7Nr05QdOcKsjtDi4U2DdSB5D7p1ky51mJ6csDIRQXtrFyPf7RRtCkXQKOl1xJl1bPNzP8WxpJGDi2LYNJpkpSm6nbKvZnfCpreFCscBGphePeyMrTlP0k2I6OB34N5ekAHd5+AaqfYvGPHTlTLFKOR6byngdp88bblo5925K0zo2CT1WXSR2uJk2BHm7avlwWQP26r6RQ=

- Traditional OOD name, not needed
- The targeted POSIX group
- The compressed and encrypted message

What’s in a name?


```
{
  "accounting": {
    "mtu": "124422,124423",
    "mti": "5a2c2e29-18e8-47d6-b413-88a24a09ae6e",
    "mta": "debug",
    "mtm": "card",
    "mtd": "sys/multitenant-delivery_default"
  },
  "connection": {
    "host": "cpu-amd-05.deac.wfu.edu",
    "ollama_url": "http://cpu-amd-05.deac.wfu.edu:63382",
    "ollama_models": "/data/ollama",
    "ollama_model": "gpt-oss:120b"
  }
}
```

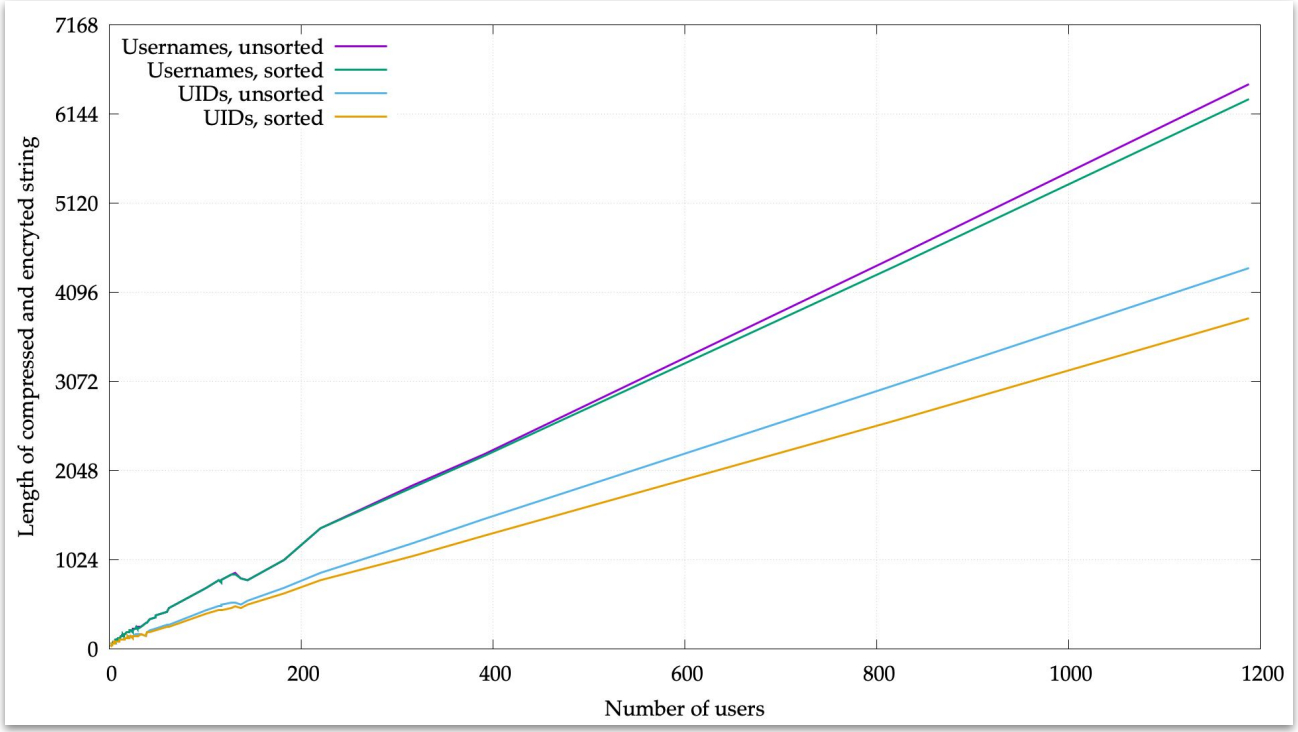
What's in a name?

```
"mtu": "124422,124423",
```

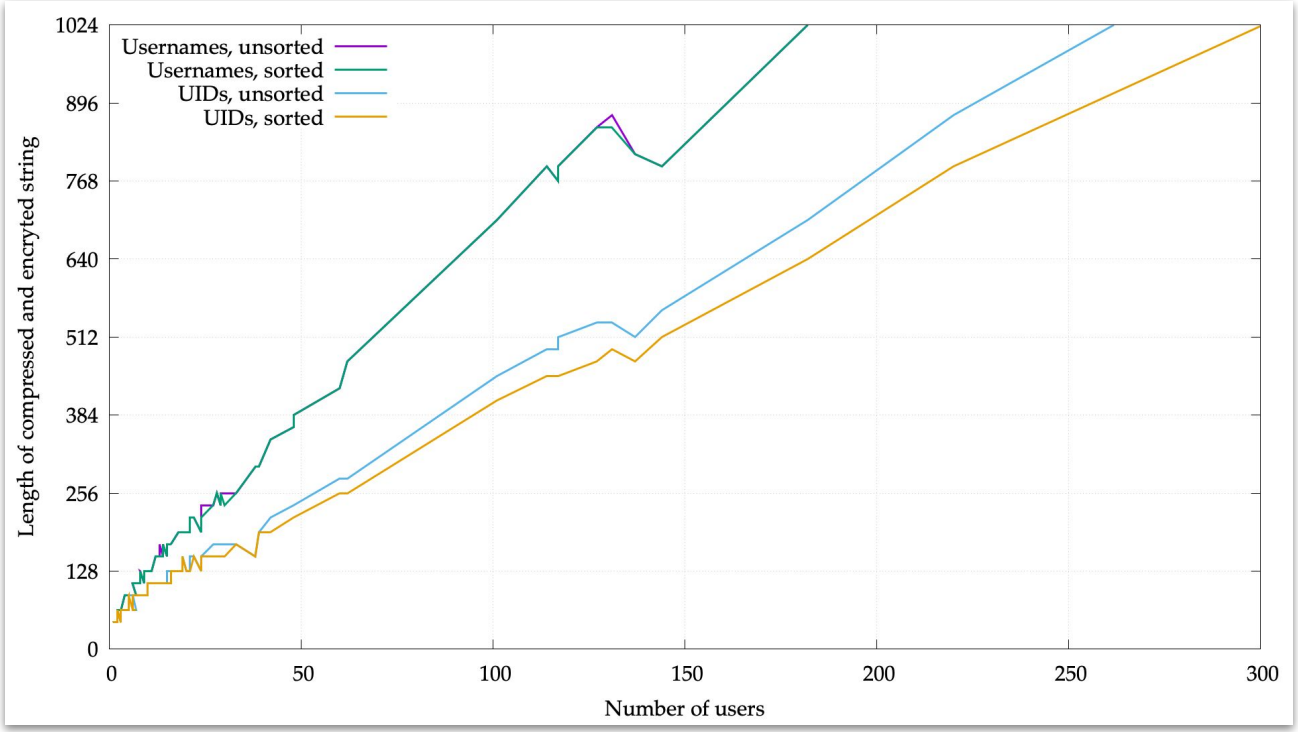
- Autogenerated
- UIDs are well behaved, sortable
- Best compression ratio

1024 character limit:
Realistically 150–200 users

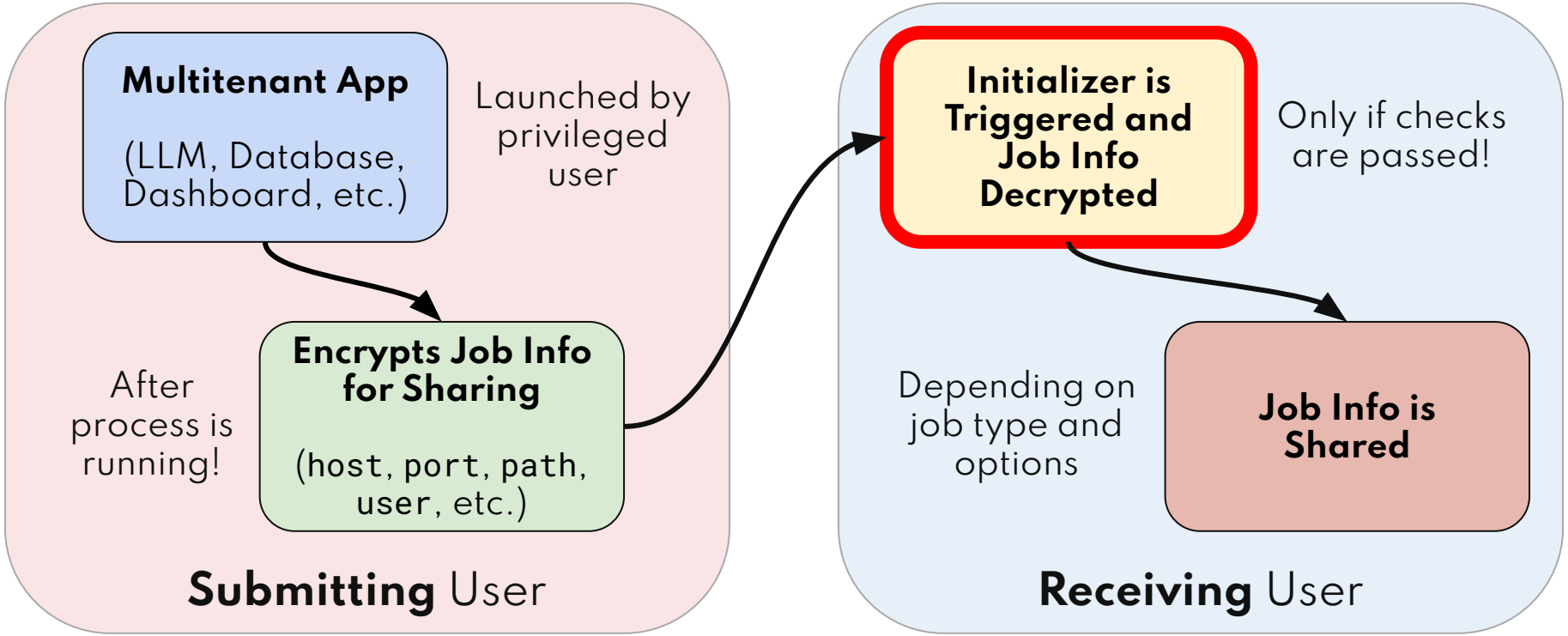
What's in a name?



What's in a name?



Overview


Multitenant App
(LLM, Database, Dashboard, etc.)
Launched by privileged user
Encrypts Job Info for Sharing
(host, port, path, user, etc.)
After process is running!
Submitting User
Initializer is Triggered and Job Info Decrypted
Only if checks are passed!
Job Info is Shared
Depending on job type and options
Receiving User

Multitenant Initializer

Initializer is Triggered and Job Info Decrypted

At the heart of the initializer, two Slurm commands:

```
sacct --state=RUNNING --format=jobidraw,user,cluster -a --wckeys=multitenant
```

IF that returns any jobs:

```
squeue --format=%i|%1024j -j <jobids>
```

Multitenant Initializer

sys/dashboard|**hpcGrp**|FTYql6YwioiGTCTaEKHY686l6ZtQLnv/m
/7Nr05QdOcKsjtDi4U2DdSB5D7p1ky51mJ6csDIRQXtrFyPf7RRtCk
XQKOl1xJl1bPNzP8WxpJGDi2LYNJpkpSm6nbKvZnfCpreFCscBGphe
PeyMrTlP0k2I6OB34N5ekAHd5+AaqfYvGPHTlTLFKOR6byngdp88bb
lo5925K0zo2CT1WXSR2uJk2BHm7avlwWQP26r6RQ=

- Traditional OOD name, not needed
- **The targeted POSIX group**
- The compressed and encrypted message

Multitenant Initializer

sys/dashboard|**hpcGrp**|FTYql6YwioiGTCTaEKHY686l6ZtQLnv/m
/7Nr05QdOcKsjtDi4U2DdSB5D7p1ky51mJ6csDIRQXtrFyPf7RRtCk
XQKOl1xJl1bPNzP8WxpJGDi2LYNJpkpSm6nbKvZnfCpreFCscBGphe
PeyMrTlP0k2I6OB34N5ekAHd5+AaqfYvGPHTlTLFKOR6byngdp88bb
lo5925K0zo2CT1WXSR2uJk2BHm7avlwWQP26r6RQ=

After initial Slurm commands, the rest of the initializer only goes into effect if “Receiving User” in POSIX group

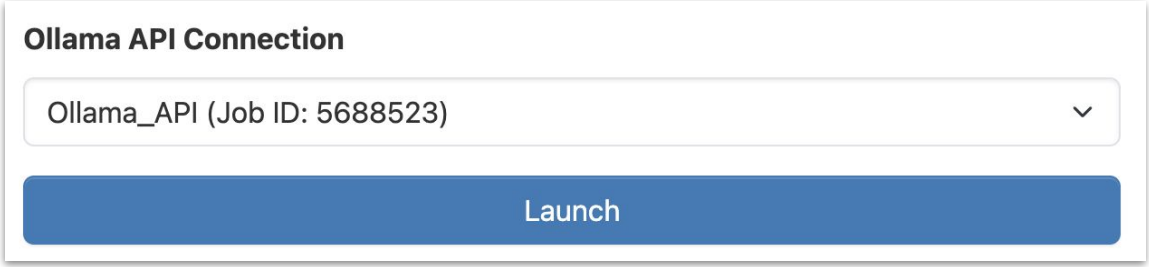
Multitenant Initializer

Once multitenant jobs have been found and are valid for user, a Ruby hash is populated with the information.

Initializer is Triggered and Job Info Decrypted

This Ruby hash is always available after initializer is run:

```
MultiTenant.specs
```



Multitenant Initializer

Once multitenant jobs have been found and are valid for user, a Ruby hash is populated with the information.

Initializer is Triggered and Job Info Decrypted

Only **two** files need to be created to have the job info appear in card:

- `${DATAROOT}/batch_connect/db/<session_id>`
- `${DATAROOT}/batch_connect/<app_name>/output/<session_id>/connection.yml`

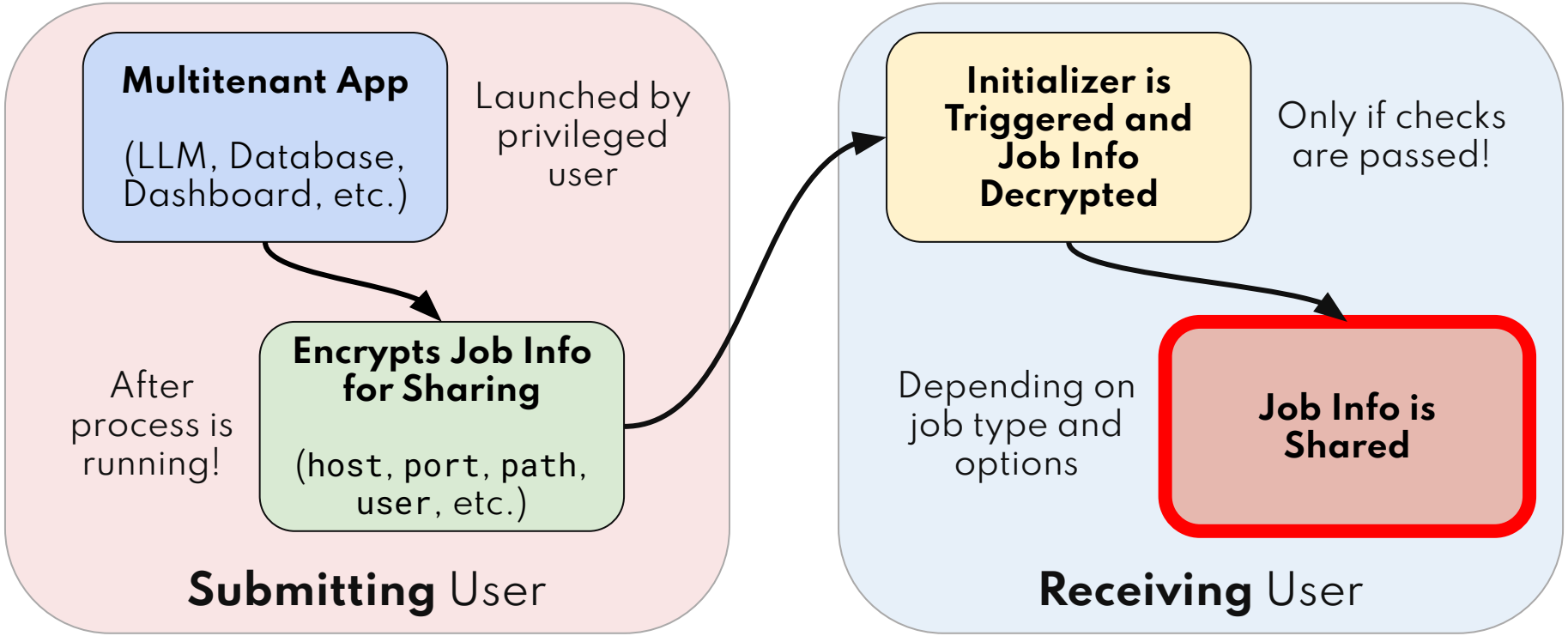
Multitenant Initializer

Designed to be **VERY LOW** impact on OOD and Slurm servers!

Initializer is Triggered and Job Info Decrypted

1. **WCKey** filter is slower but has much lower impact than `grep`
2. **Slurm job name** is only captured on valid Job IDs
3. **Encrypted string** is only decrypted and uncompressed when all user requirements are met
4. **`db` and `connection.yml` files** only get created if four different checks are passed!

Overview


Multitenant App
(LLM, Database, Dashboard, etc.)
Launched by privileged user
Encrypts Job Info for Sharing
(host, port, path, user, etc.)
After process is running!
Submitting User
Initializer is Triggered and Job Info Decrypted
Only if checks are passed!
Job Info is Shared
Depending on job type and options
Receiving User

Content Delivery

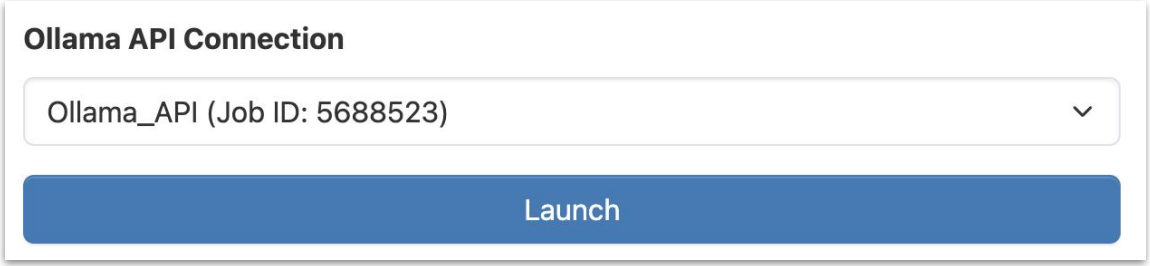
If **FORM VIEW ONLY** was selected:

```
ollama_interface:
  label: "Ollama API Connection"
  required: false
  widget: select
  options:
    <%- MultiTenant.specs.each do |key, value| %>
    - [
        '<%= "#{value["accounting"]["mta"]} (Job ID: #{key})" %>',
        '<%= value["connection"]['ollama_url'] %>'
      ]
    <%- end -%>
  cacheable: false
```

Job Info is Shared

Content Delivery

If **FORM VIEW ONLY** was selected:



Job Info is Shared

Content Delivery

If **CARD + FORM VIEW** was selected:

We now have:

- `db` file that creates the card in the dashboard
- `connection.yml` that has all of the info to be displayed in the card

Job Info is Shared

Now we just need a `view.html.erb`

Content Delivery

Delivery Apps

"Fake" apps that only have the `view.html.erb`:

- `delivery_default`
- `delivery_debug`
- `delivery_readfile`

Job Info is Shared

Content Delivery

delivery_default

```
<%- if mt_appname == "Ollama_API" -%>

<strong>Ollama API URL:</strong><pre><code><%= ollama_url.to_s
%></code></pre>

The Ollama API URL shown above has been automatically set in the
<code>$OLLAMA_HOST</code> environment variable.

<form action="sys/multitenant-jupyter" method="get" target="_self">
  <button class="btn btn-primary" type="submit">
    <i class="fa fa-eye"></i> Launch Jupyter
  </button>
</form>

<%- end -%>
```

Job Info is Shared

Content Delivery

`delivery_default`

Ollama_API from hpcfaculty (5688517) 1 node | 64 cores | Running

Host: gpu-a100-03.deac.wfu.edu

Cancel

Created at: 2025-09-29 23:25:00 EDT

Time Remaining: 57 minutes

Session ID: bf6007d3-bd48-4206-9ff1-e33d2e0fc677

Problems with this session? Submit support ticket

Ollama API URL:

```
http://gpu-a100-03.deac.wfu.edu:44651
```

The Ollama API URL shown above has been automatically set in the `$OLLAMA_HOST` environment variable.

For Jupyter notebooks: Select the Ollama API from the list in the form. Once in the Notebook, list the models available like this:

```
import ollama
for m in ollama.list().models:
  print(m.model)
```

Launch Jupyter | Launch Gradio

Job Info is Shared

Content Delivery

delivery_debug

```
<div style="margin-bottom: 4px;">
    <h5>Multitenant Debug</h5>
    <pre>
    <%= JSON.pretty_generate(MultiTenant.specs["#{jobid}"]) %>
    </pre>
</div>
```

Job Info is Shared

Content Delivery

`delivery_debug`

Multitenant Debug


```
{
  "info": {
    "user": "hpcfaculty",
    "cluster": "deac",
    "db": "/home/hpcstudent/ondemand/data/sys/dashboard/batch_connect/db/5284d73c-ea93-4b3f-ba71-9fa8d43
    "output": "/home/hpcstudent/ondemand/data/sys/dashboard/batch_connect/sys/multitenant-delivery_debug,
    "message": "sys/dashboard|hpcGrp|BnZTNva0kSe7y6/9o8RLhQXq6rQz3XBY6NLExCkx75nBBaYr+qY37wOdqNfiso925W9
    "message_size": 237
  },
  "accounting": {
    "mtu": "124422,124423",
    "mti": "5284d73c-ea93-4b3f-ba71-9fa8d439d06a",
    "mta": "debug",
    "mtm": "card",
    "mtd": "sys/multitenant-delivery_debug"
  },
  "connection": {
    "host": "cpu-amd-05.deac.wfu.edu",
    "port": "63087",
    "jobid": "5688522",
    "mt_appname": "debug"
  }
}
```


Job Info is Shared

```
{
  "info": {
    "user": "hpcfaculty",
    "cluster": "deac",
    "db":
"/home/hpcstudent/ondemand/data/sys/dashboard/batch_connect/db/5284d73c-ea93-4b3f-ba71-9fa8d439d06
a",
    "output":
"batch_connect/sys/multitenant-delivery_debug/output/5284d73c-ea93-4b3f-ba71-9fa8d439d06a",
    "message":
"sys/dashboard|hpcGrp|BnZTNva0kSe7y6/9o8RLhQXq6rQz3XBY6NLExCkx75nBBaYr+qY37wOdqNfiso925W9qSqXDi7af
g5jCir3F52JOFXp9TXyyQvNbHO8QwR+3Hi1rOffp06Q5kUMjlAnH36uy6rmwCfdpGmc+0cuIKqqoobvWFNFVDA3y1nBN1HYBEk
UOK3WfzNzP25p8gwcrHT1CbbNlwTIPBo7dqG7gpg==",
    "message_size": 237
  },
  "accounting": {
    "mtu": "124422,124423",
    "mti": "5284d73c-ea93-4b3f-ba71-9fa8d439d06a",
    "mta": "debug",
    "mtm": "card",
    "mtd": "sys/multitenant-delivery_debug"
  },
  "connection": {
    "host": "cpu-amd-05.deac.wfu.edu",
    "port": "63087",
    "jobid": "5688522",
    "mt_appname": "debug"
  }
}
```

```
{
  "info": {
    "user": "hpcfaculty",
    "cluster": "deac",
    "db":
"/home/hpcstudent/ondemand/data/sys/dashboard/batch_connect/db/5284d73c-ea93-4b3f-ba71-9fa8d439d06
a",
    "output":
"batch_connect/sys/multitenant-delivery_debug/output/5284d73c-ea93-4b3f-ba71-9fa8d439d06a",
    "message":
"sys/dashboard|hpcGrp|BnZTNva0kSe7y6/9o8RLhQXq6rQz3XBY6NLExCkx75nBBaYr+qY37wOdqNfiso925W9qSqXDi7af
g5jCir3F52JOFXp9TXyyQvNbHO8QwR+3Hi1rOffp06Q5kUMjlAnH36uy6rmwCfdpGmc+0cuIKqqoobvWFNFVDA3y1nBN1HYBEk
UOK3WfzNzP25p8gwcrHT1CbbNlwTIPBo7dqG7gpg==",
    "message_size": 237
  },
  "accounting": {
    "mtu": "124422,124423",
    "mti": "5284d73c-ea93-4b3f-ba71-9fa8d439d06a",
    "mta": "debug",
    "mtm": "card",
    "mtd": "sys/multitenant-delivery_debug"
  },
  "connection": {
    "host": "cpu-amd-05.deac.wfu.edu",
    "port": "63087",
    "jobid": "5688522",
    "mt_appname": "debug"
  }
}
```


Original Message

Added in Initializer


```
{
  "info": {
    "user": "hpcfaculty",
    "cluster": "deac",
    "db":
"/home/hpcstudent/ondemand/data/sys/dashboard/batch_connect/db/5284d73c-ea93-4b3f-ba71-9fa8d439d06
a",
    "output":
"batch_connect/sys/multitenant-delivery_debug/output/5284d73c-ea93-4b3f-ba71-9fa8d439d06a",
    "message":
"sys/dashboard|hpcGrp|BnZTNva0kSe7y6/9o8RLhQXq6rQz3XBY6NLExCkx75nBBaYr+qY37wOdqNfiso925W9qSqXDi7af
g5jCir3F52JOFXp9TXyyQvNbHO8QwR+3Hi1rOffp06Q5kUMjlAnH36uy6rmwCfdpGmc+0cuIKqqoobvWFNFVDA3y1nBN1HYBEk
UOK3WfzNzP25p8gwcrHT1CbbNlwTIPBo7dqG7gpg==",
    "message_size": 237
  },
  "accounting": {
    "mtu": "124422,124423",
    "mti": "5284d73c-ea93-4b3f-ba71-9fa8d439d06a",
    "mta": "debug",
    "mtm": "card",
    "mtd": "sys/multitenant-delivery_debug"
  },
  "connection": {
    "host": "cpu-amd-05.deac.wfu.edu",
    "port": "63087",
    "jobid": "5688522",
    "mt_appname": "debug"
  }
}
```


Original Message

Content Delivery

delivery_readfile

```
<%= File.read(content_path) %>
```

Job Info is Shared

Security/Safeguarding

- **Encrypting connection details in Slurm name**
  - Infinite variations on format and credentials
  - OOD admin can change at will (initializer/`after.sh.erb`)
  - Jobs are ephemeral in nature!

Security/Safeguarding

- **Who can see the code?**
  - Only privileged user can see `after.sh.erb`
    - Trusted PI
    - OOD or HPC admin
  - Generated `connection.yml` files are business as usual
  - Only the OOD admin can see initializer!

Security/Safeguarding

- **POSIX Groups for controlling app access**
  - Good for first approximation
  - Can be clunky if you rely on AD

Security/Safeguarding

- **Slurm WCKeys for controlling app access**
  - Take effect immediately
  - Slurm/OOD admin friendly

Security/Safeguarding

*WCKeys (OPTIONAL)

- Add to `slurm.conf`:

  `AccountingStorageEnforce=wckeys,...`

  `TrackWCKey=yes`

- Add to `slurmdbd.conf`:

  `TrackWCKey=yes`

Security/Safeguarding

WCKeys (OPTIONAL)

```
$ sacctmgr add user hpcfaculty wckey=multitenant
```

Security/Safeguarding

WCKeys (OPTIONAL)

```
$ sacctmgr del user hpcfaculty wckey=multitenant
```

Failed to submit session with the following error:

```
sbatch: error: Batch job submission failed: Invalid wckey specification
```

- If this job failed to submit because of an invalid job name please ask your administrator to configure OnDemand
  OOD_JOB_NAME_ILLEGAL_CHARS.
- The Multitenant LLM session data for this session can be accessed under the staged root directory.

Conclusions

- **We can (carefully) share job info between users**
  - Content delivered through forms or cards
  - Shared resources work as expected
- **Leveraged OOD interaction with Slurm**
  - No custom JS, no grepping, no file-updating
  - Initializer and some Bash scripting
- **Minimal modification to Slurm config**
  - WCKeys add security and minimize overhead

Future Work

- **Slurm comment**?
- `sacct` bug: 256 character limit for job name

- Kubernetes and other schedulers
- OOD workflows with pre-filled forms or linked configs

Use Cases

We want to:

1. Offer an LLM software stack that users can access from within the cluster
2. Offer a database software stack that select users can access from within the cluster
3. Enable PIs to share dashboards and other web services within their research groups or departments
4. **??? (Bonus)**

THANKS!

anderss@wfu.edu
https://hpc.wfu.edu

WAKE
FOREST
UNIVERSITY