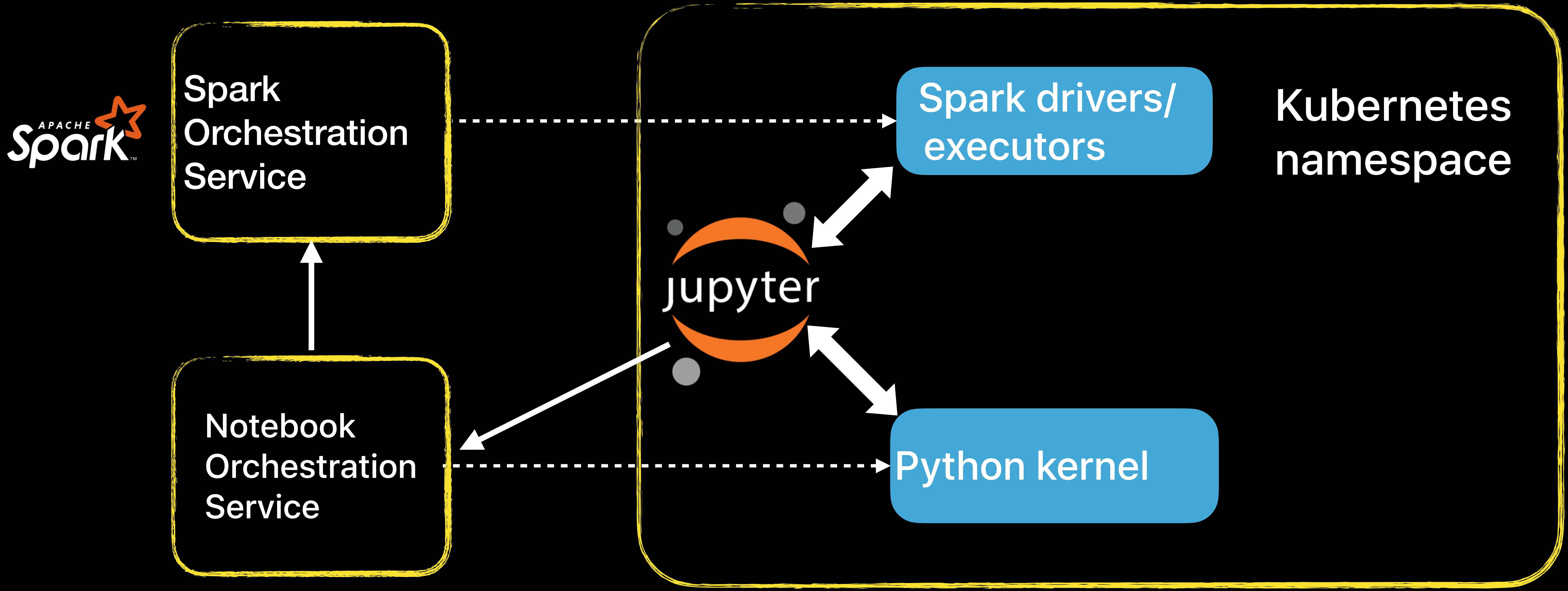
Data Science Environments In Seconds: Scaling Jupyter Notebooks in Kubernetes

Jialin Zhang

AppDeveloperCon | Apple | Nov 12, 2024


Spark Orchestration Service
Spark drivers/ executors
Kubernetes namespace
jupyter
Notebook Orchestration Service
Python kernel

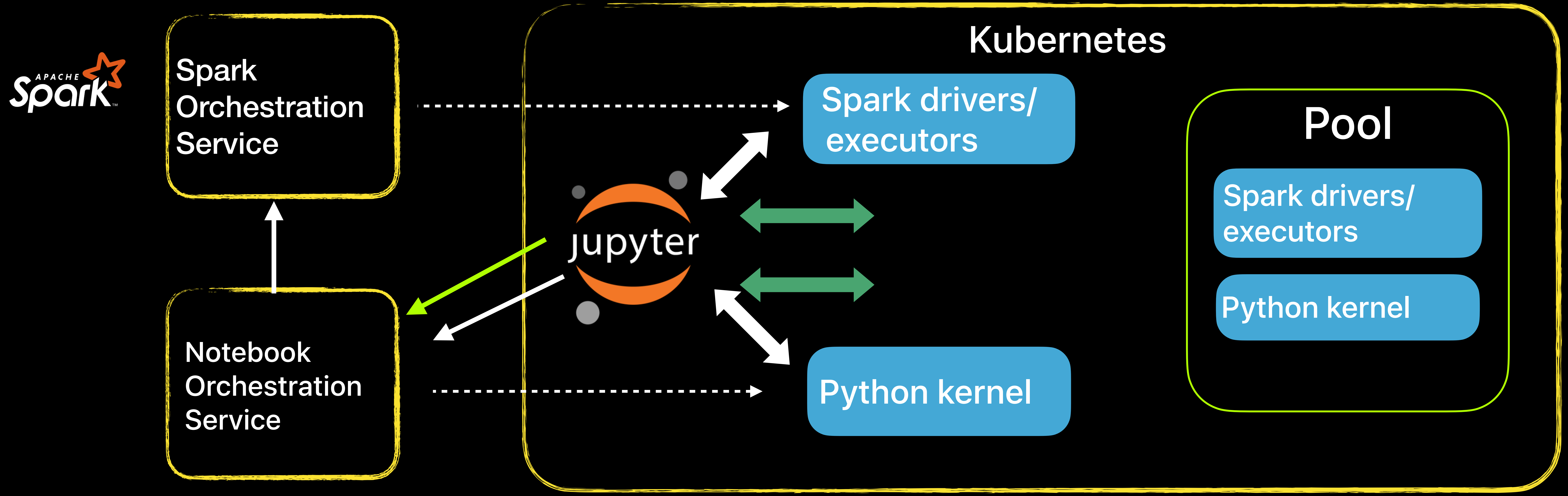
Challenge

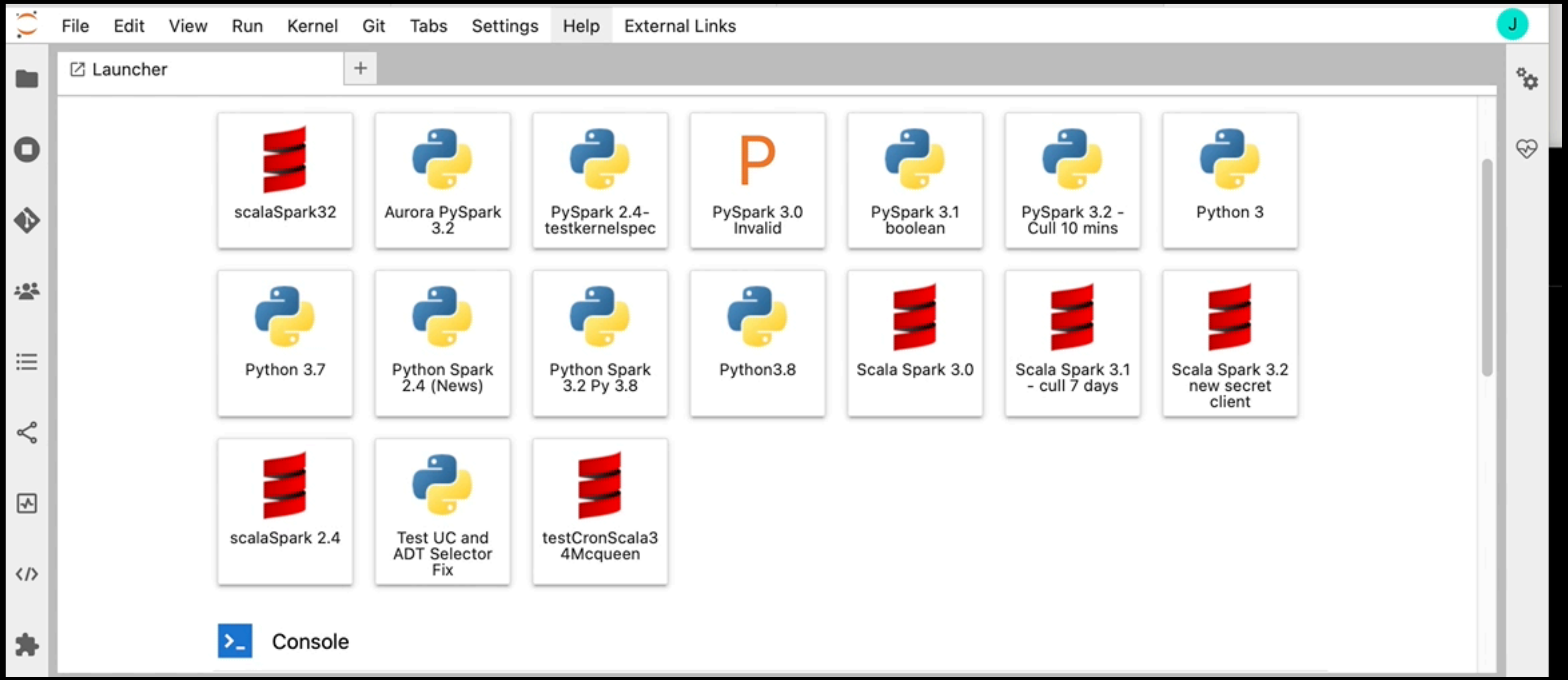
Kubernetes Pod Kernel Launch takes minutes

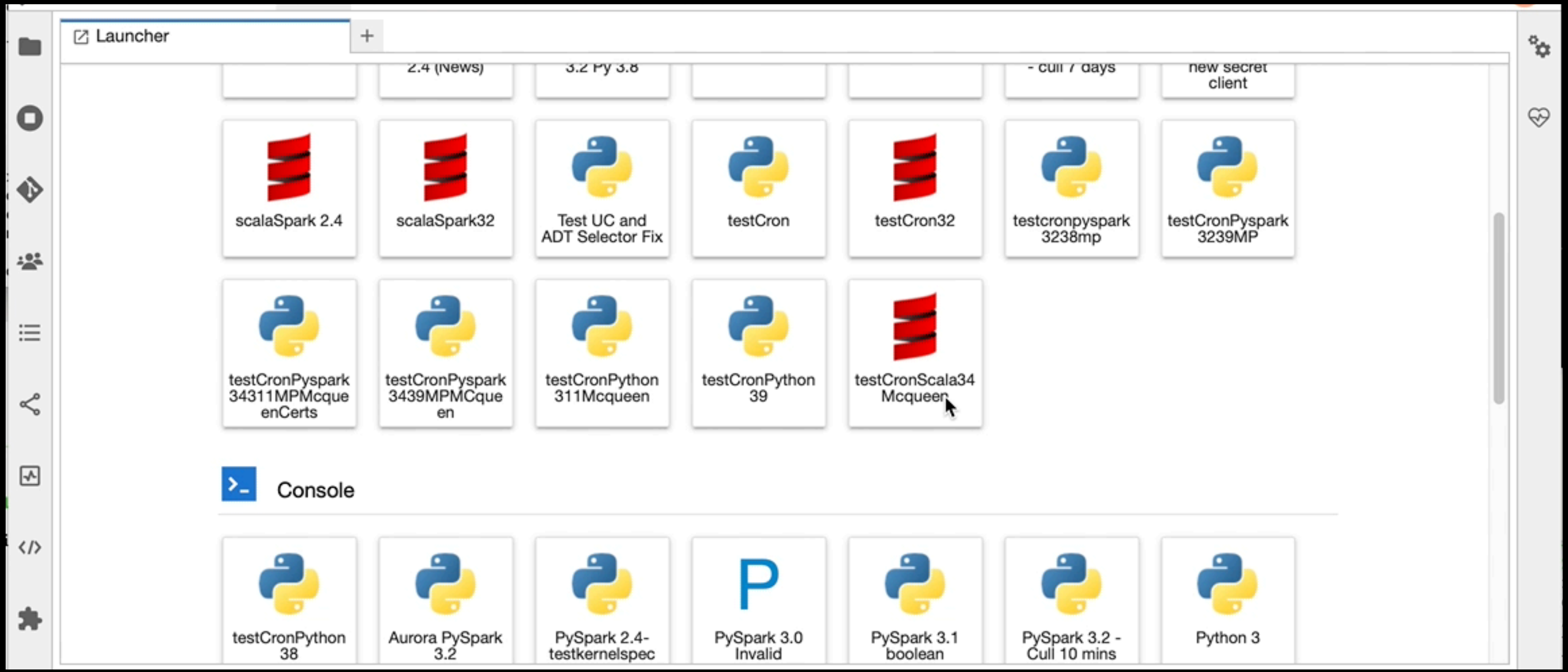
Solution

Pre-warm kernel in pool

If user wants kernel available in pool, directly pick up from pool

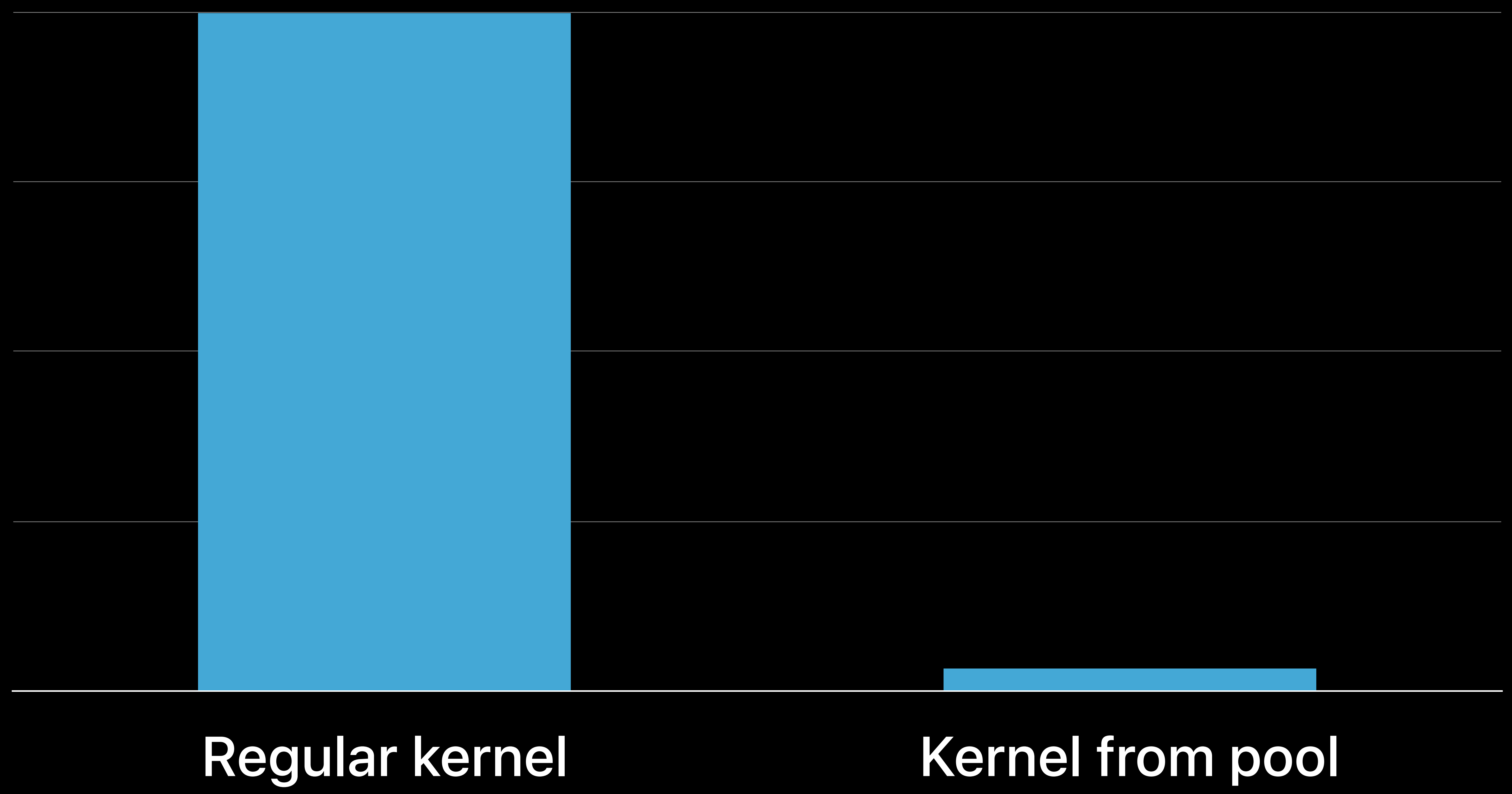



File
Edit
View
Run
Kernel
Git
Tabs
Settings
Help
External Links
J
Launcher
scalaSpark32
Aurora PySpark 3.2
PySpark 2.4-testkernelspec
PySpark 3.0 Invalid
PySpark 3.1 boolean
PySpark 3.2 - Cull 10 mins
Python 3
Python 3.7
Python Spark 2.4 (News)
Python Spark 3.2 Py 3.8
Python3.8
Scala Spark 3.0
Scala Spark 3.1 - cull 7 days
Scala Spark 3.2 new secret client
scalaSpark 2.4
Test UC and ADT Selector Fix
testCronScala34Mcqueen
Console


Launcher
2.4 (News)
3.2 Py 3.8
- cull 7 days
new secret client
scalaSpark 2.4
scalaSpark32
Test UC and ADT Selector Fix
testCron
testCron32
testcronpyspark3238mp
testCronPyspark3239MP
testCronPyspark34311MPMcqueenCerts
testCronPyspark3439MPMcqueen
testCronPython311Mcqueen
testCronPython39
testCronScala34Mcqueen
Console
testCronPython38
Aurora PySpark 3.2
PySpark 2.4-testkernelspec
PySpark 3.0 Invalid
PySpark 3.1 boolean
PySpark 3.2 - Cull 10 mins
Python 3

Improved Productivity

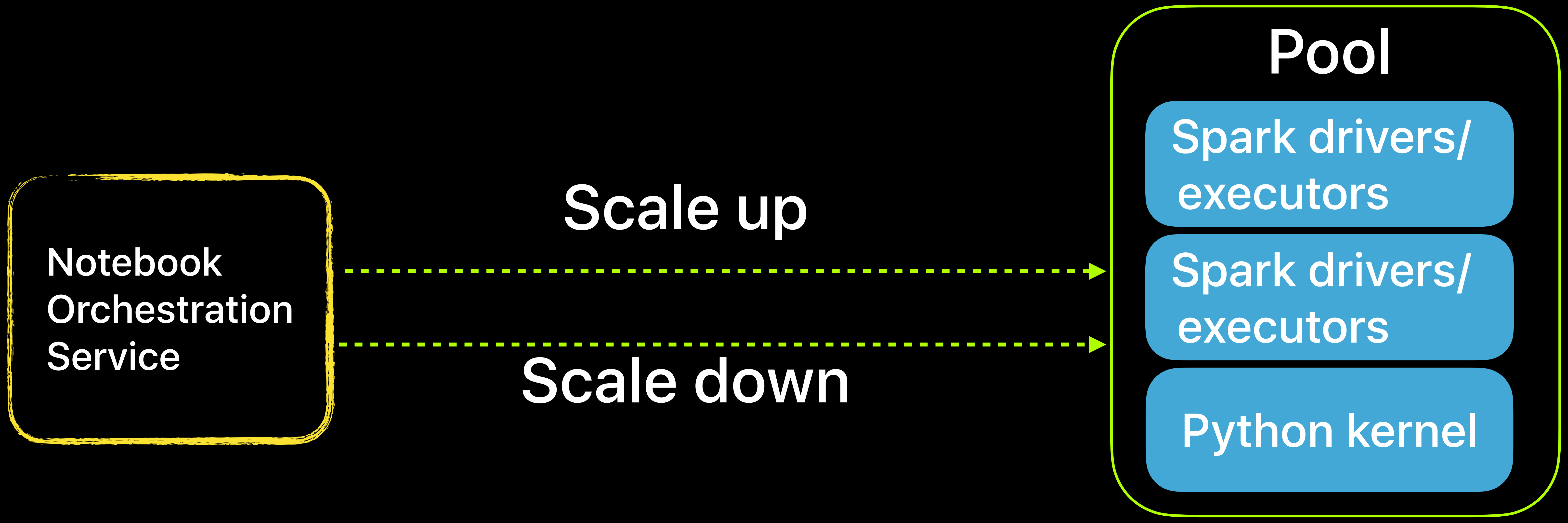
Reduce launch latency by 96%, from minutes to seconds



Architecture + Design

Resource driven dynamic pool management

- Periodically
- Scale up/down kernels: Pool + existing usage < 95% quota usage
- Pool max size <= 10% of namespace quota
- Prediction -> rank by launch frequency



Architecture + Design

Cost Savings

- Large Namespace
- Actively detect kernels in pools that were running for a long time and tear them down
- Kernel configuration with small initial resource footprint

Pool per Namespace

- Auto-detect eligible namespaces dynamically

Areas to Explore

Improving prediction

Time triggered scaling —> Watcher triggered scaling

More namespaces with lower resource quota