



KubeCon

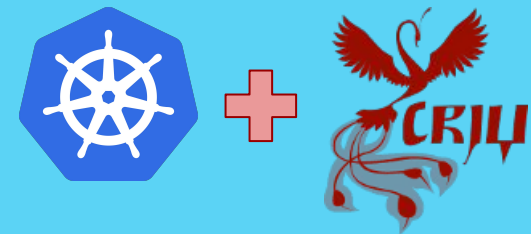


CloudNativeCon

Europe 2024



Enabling Coordinated Checkpointing for Distributed HPC Applications



Radostin Stoyanov - PhD Student @ Scientific Computing Group

Adrian Reber - Senior Principal Software Engineer

Supervisor: Prof. Wes Armour

Paris, 2024, March 22



Agenda



- Background
- Integrations
- Use Cases
- Migration Demo
- Coordinated Checkpointing

The background features a composition of overlapping geometric shapes in various shades of blue and purple. A large, dark purple shape dominates the right side, while a medium blue shape is positioned in the upper left. A lighter blue shape is visible in the lower right, partially overlapping the dark purple one. The overall aesthetic is modern and minimalist.

Background



KubeCon



CloudNativeCon

Europe 2024

Checkpoint/Restore in Userspace CRIU

Integrations



KubeCon



CloudNativeCon

Europe 2024

Multiple Integrations Exist



KubeCon



CloudNativeCon

Europe 2024

OpenVZ



KubeCon



CloudNativeCon

Europe 2024

Borg



KubeCon



CloudNativeCon

Europe 2024

LXC / LXD / Incus



KubeCon



CloudNativeCon

Europe 2024

Docker



KubeCon



CloudNativeCon

Europe 2024

Podman



KubeCon



CloudNativeCon

Europe 2024

CRI-O

Forensic Container Checkpointing

- <https://github.com/kubernetes/enhancements/pull/1990> (alpha 1.25)
- <https://github.com/kubernetes/enhancements/pull/3264> (alpha 1.25)
- <https://github.com/kubernetes/kubernetes/pull/104907> (alpha 1.25)
- <https://kubernetes.io/blog/2022/12/05/forensic-container-checkpointing-alpha/>
- <https://kubernetes.io/blog/2023/03/10/forensic-container-analysis/>
- <https://www.opensourcerers.org/2023/09/11/forensic-container-checkpointing-in-openshift>
- <https://github.com/kubernetes/enhancements/pull/4288> (beta 1.30)
- <https://github.com/kubernetes/kubernetes/pull/123215> (beta 1.30)

Use Cases

Migration Demo

Checkpointing Distributed HPC Applications

Checkpointing Distributed HPC Applications

Problem definition and research questions

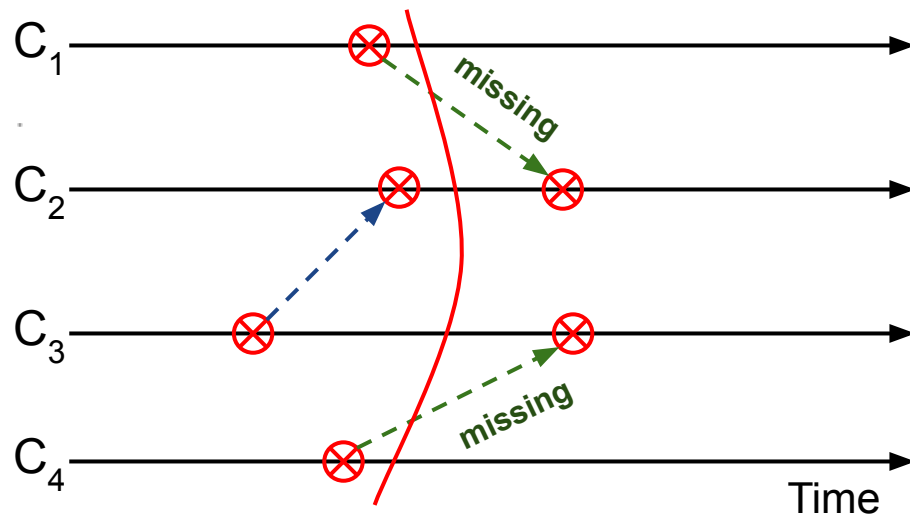
- CRIU was designed for checkpoint/restore of individual process tree, rather than distributed applications
- How to extend CRIU with support for distributed applications?
- How to enable checkpoint coordination within Kubernetes?

Related work

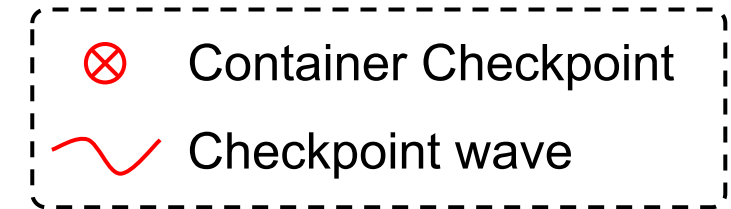
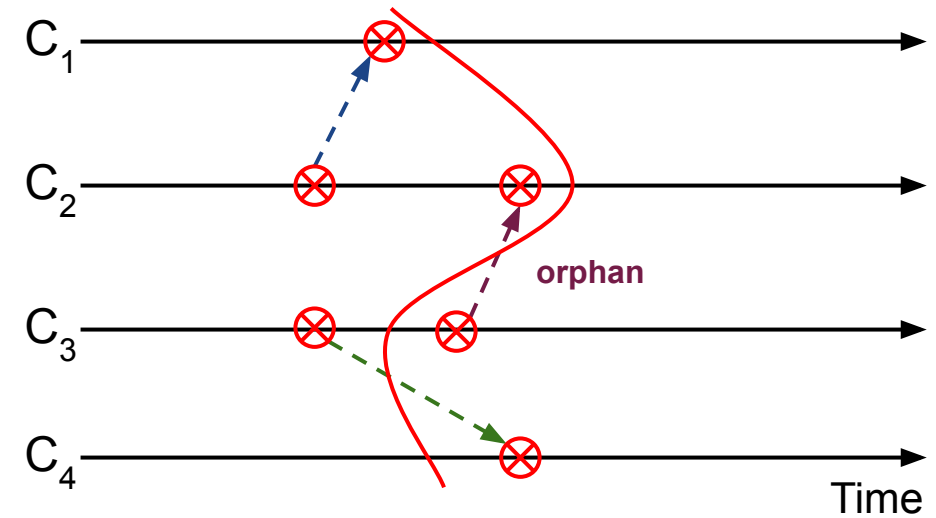
- DMTCP (Distributed MultiThreaded CheckPointing)
- Apache Flink

Coordinated Checkpointing

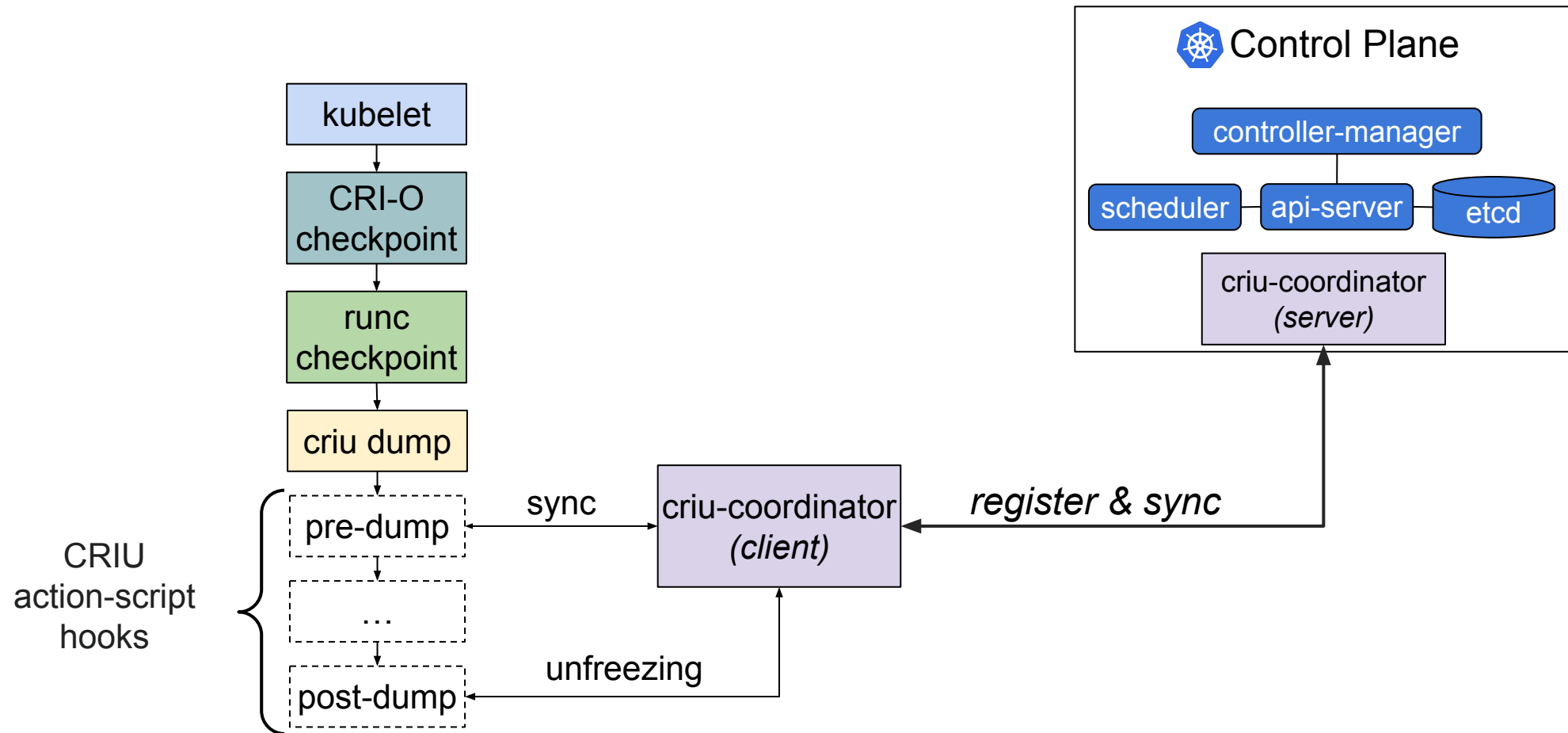
Consistent checkpoint



Inconsistent checkpoint

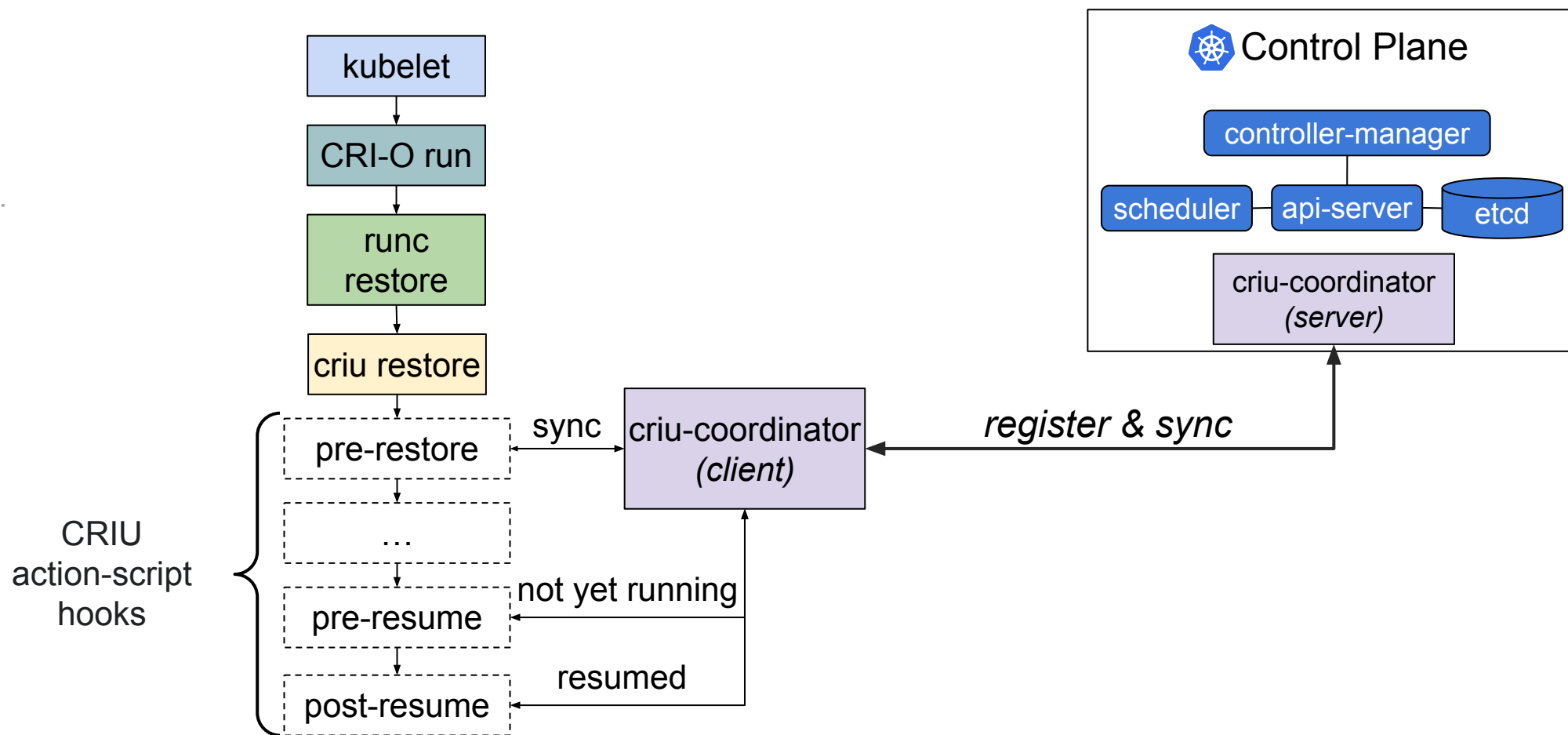


CRIU Coordination Protocol - Checkpointing





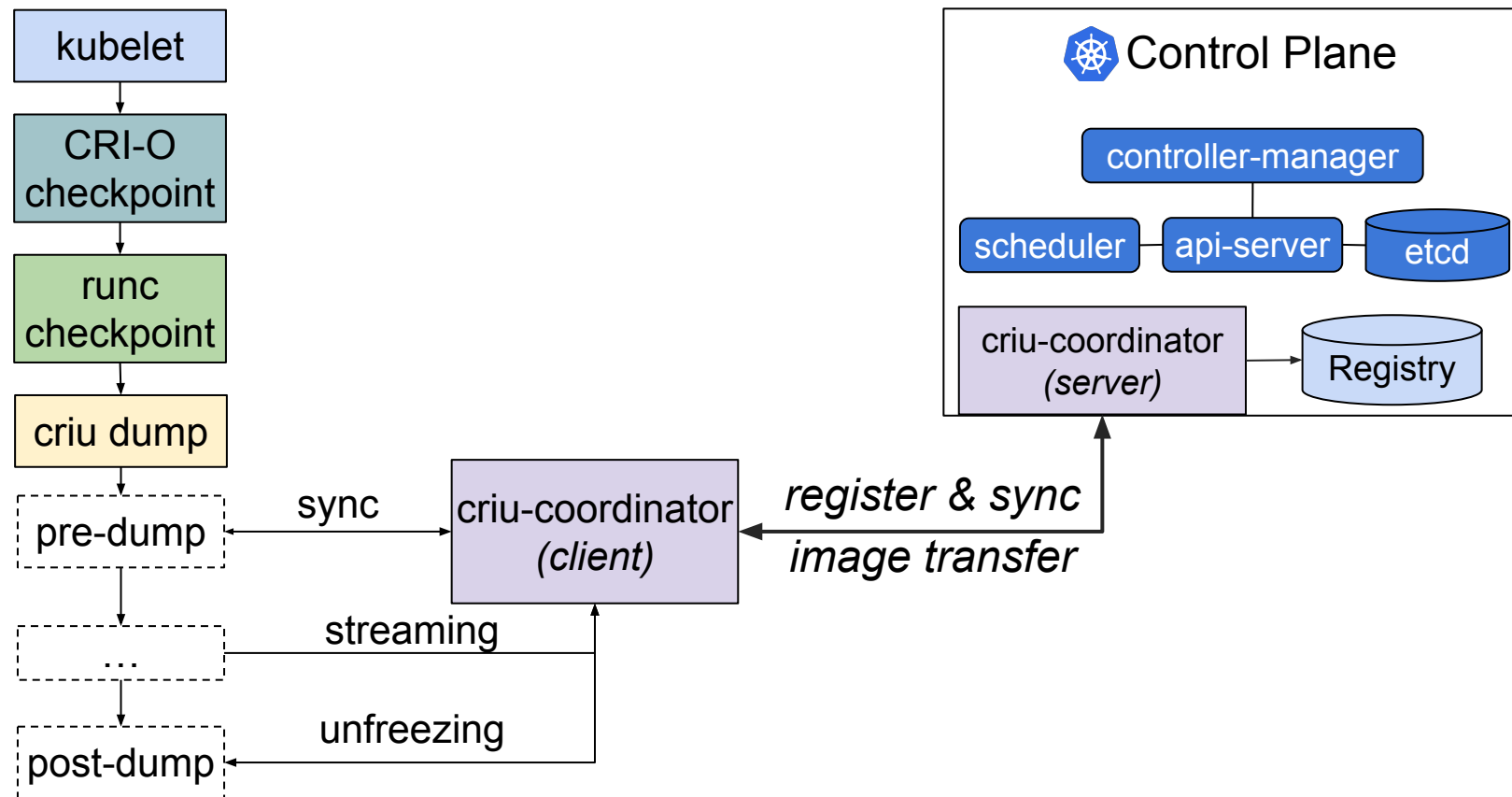
CRIU Coordination Protocol - Restore



Coordinated Checkpointing Demo

<pre> 08.834576) my: Internal View 2 08.834578) Writing Image Memory Overlay 11 08.834609) Running post-fpga scripts 08.834624) [./run/fpga/wr-wr-fpga.sh] 08.837029) Initializing table info 2 08.837044) Initializing T104 table 2 08.837074) Writing vram 08.837449) Script finished successfully CrashInfo: [./run/subprocess4-demo] 0 20: 8 </pre>		<pre> 194 - [Client-81] [in] Checking connection w/ Client-81 195 - [Client-81] [in] Client-81 connected 196 - [Client-81] [in] Client is ready 197 - [Client-81] [in] Wait for all dependencies to be ready 198 - [Client-81] [in] Checking readiness w/ Client-81 199 - [Client-81] [in] Dependency Client-81 is ready 200 - [Client-81] [in] Sending W0 201 - [in] New Client connected: 19.208.17.31:30999 202 - [in] Initialize Client 19, action and dependencies 203 - [Client-81] [in] 19: Client-81 204 - [Client-81] [in] ACTION: post-fpga 205 - [Client-81] [in] BROADCAST: Client-81 206 - [Client-81] [in] Wait for all dependencies to be ready 207 - [Client-81] [in] Checking local dependencies: Client-81 208 - [Client-81] [in] Sending W0 209 - [Client-81] [in] Client-81 connected 210 - [Client-81] [in] Client is ready 211 - [Client-81] [in] Wait for all dependencies to be ready 212 - [Client-81] [in] Checking readiness w/ Client-81 213 - [Client-81] [in] Dependency Client-81 is ready 214 - [Client-81] [in] Sending W0 215 - [in] New Client connected: 19.208.17.31:30999 216 - [in] Initialize Client 19, action and dependencies 217 - [Client-81] [in] 19: Client-81 218 - [Client-81] [in] ACTION: post-fpga 219 - [Client-81] [in] BROADCAST: Client-81 220 - [Client-81] [in] Wait for all dependencies to be ready 221 - [Client-81] [in] Checking local dependencies: Client-81 222 - [Client-81] [in] Sending W0 223 - [Client-81] [in] Client disconnected </pre>
<pre> 08.839627) my: Internal View 3 08.839629) Writing Image Memory Overlay 11 08.839630) Running post-fpga scripts 08.839632) [./run/fpga/wr-wr-fpga.sh] 08.839643) Initializing table info 3 08.839673) Initializing T118 table 3 08.839687) Writing vram 08.839841) Script finished successfully CrashInfo: [./run/subprocess4-demo] 0 20: 8 </pre>		<pre> 194 - [Client-81] [in] Checking connection w/ Client-81 195 - [Client-81] [in] Client-81 connected 196 - [Client-81] [in] Client is ready 197 - [Client-81] [in] Wait for all dependencies to be ready 198 - [Client-81] [in] Checking readiness w/ Client-81 199 - [Client-81] [in] Dependency Client-81 is ready 200 - [Client-81] [in] Sending W0 201 - [in] New Client connected: 19.208.17.31:30999 202 - [in] Initialize Client 19, action and dependencies 203 - [Client-81] [in] 19: Client-81 204 - [Client-81] [in] ACTION: post-fpga 205 - [Client-81] [in] BROADCAST: Client-81 206 - [Client-81] [in] Wait for all dependencies to be ready 207 - [Client-81] [in] Checking local dependencies: Client-81 208 - [Client-81] [in] Sending W0 209 - [Client-81] [in] Client-81 connected 210 - [Client-81] [in] Client is ready 211 - [Client-81] [in] Wait for all dependencies to be ready 212 - [Client-81] [in] Checking readiness w/ Client-81 213 - [Client-81] [in] Dependency Client-81 is ready 214 - [Client-81] [in] Sending W0 215 - [in] New Client connected: 19.208.17.31:30999 216 - [in] Initialize Client 19, action and dependencies 217 - [Client-81] [in] 19: Client-81 218 - [Client-81] [in] ACTION: post-fpga 219 - [Client-81] [in] BROADCAST: Client-81 220 - [Client-81] [in] Wait for all dependencies to be ready 221 - [Client-81] [in] Checking local dependencies: Client-81 222 - [Client-81] [in] Sending W0 223 - [Client-81] [in] Client disconnected </pre>
<pre> 08.841700) my: Internal View 4 08.841702) Writing Image Memory Overlay 11 08.841703) Running post-fpga scripts 08.841705) [./run/fpga/wr-wr-fpga.sh] 08.841716) Initializing table info 4 08.841746) Initializing T118 table 4 08.841760) Writing vram 08.841914) Script finished successfully CrashInfo: [./run/subprocess4-demo] 0 20: 8 </pre>		<pre> 194 - [Client-81] [in] Checking connection w/ Client-81 195 - [Client-81] [in] Client-81 connected 196 - [Client-81] [in] Client is ready 197 - [Client-81] [in] Wait for all dependencies to be ready 198 - [Client-81] [in] Checking readiness w/ Client-81 199 - [Client-81] [in] Dependency Client-81 is ready 200 - [Client-81] [in] Sending W0 201 - [in] New Client connected: 19.208.17.31:30999 202 - [in] Initialize Client 19, action and dependencies 203 - [Client-81] [in] 19: Client-81 204 - [Client-81] [in] ACTION: post-fpga 205 - [Client-81] [in] BROADCAST: Client-81 206 - [Client-81] [in] Wait for all dependencies to be ready 207 - [Client-81] [in] Checking local dependencies: Client-81 208 - [Client-81] [in] Sending W0 209 - [Client-81] [in] Client-81 connected 210 - [Client-81] [in] Client is ready 211 - [Client-81] [in] Wait for all dependencies to be ready 212 - [Client-81] [in] Checking readiness w/ Client-81 213 - [Client-81] [in] Dependency Client-81 is ready 214 - [Client-81] [in] Sending W0 215 - [in] New Client connected: 19.208.17.31:30999 216 - [in] Initialize Client 19, action and dependencies 217 - [Client-81] [in] 19: Client-81 218 - [Client-81] [in] ACTION: post-fpga 219 - [Client-81] [in] BROADCAST: Client-81 220 - [Client-81] [in] Wait for all dependencies to be ready 221 - [Client-81] [in] Checking local dependencies: Client-81 222 - [Client-81] [in] Sending W0 223 - [Client-81] [in] Client disconnected </pre>

CRIU Coordination Protocol - Checkpoint Streaming





KubeCon



CloudNativeCon

Europe 2024

Future work

- Adding support for **criu-coordinator** in CRI-O & containerd
- Automating checkpoint dependency detection in Kubernetes
- Integrating checkpoint/restore with Kubernetes objects (e.g., Deployments, StatefulSets, etc.)



KubeCon



CloudNativeCon

Europe 2024

