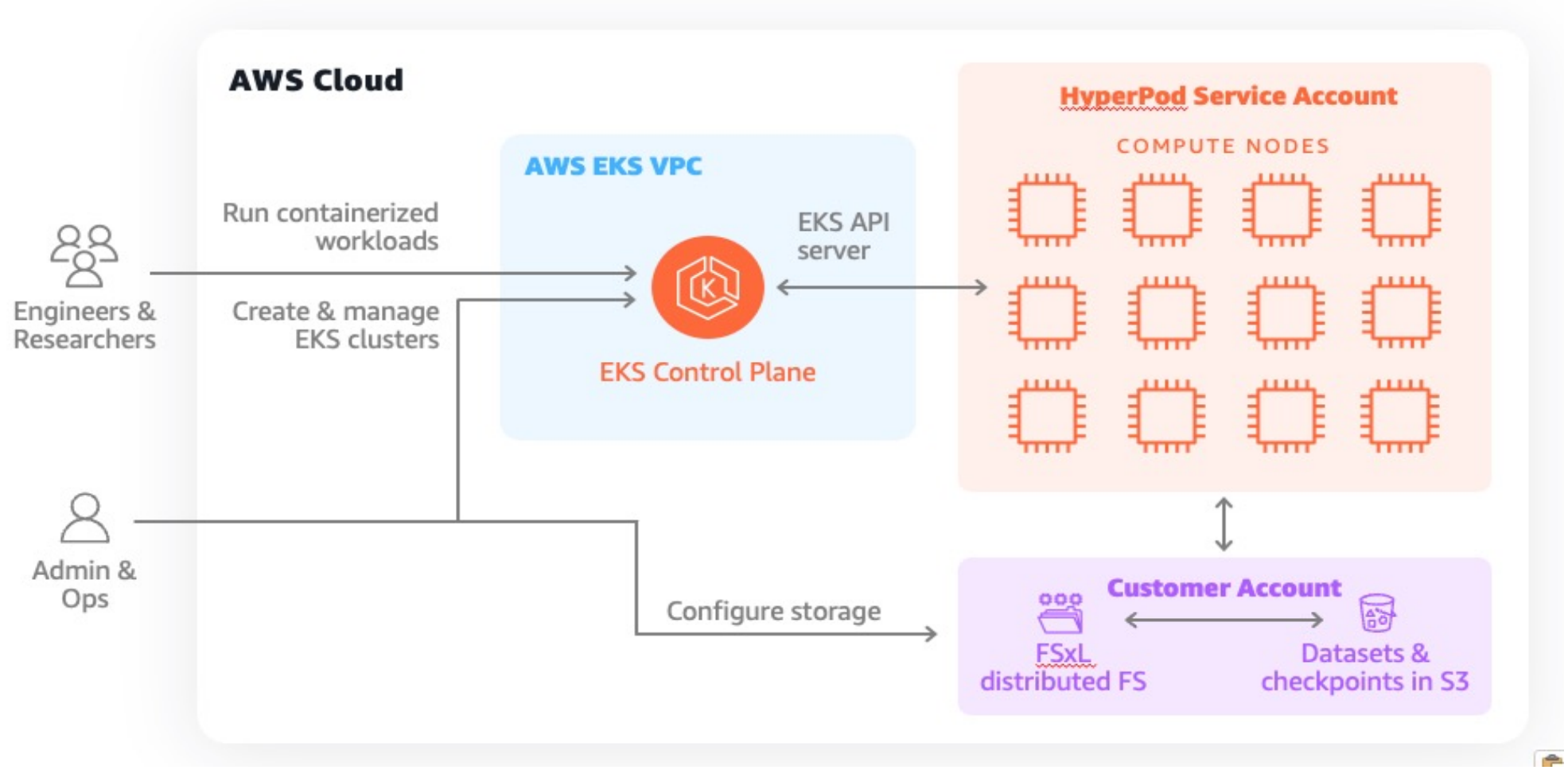


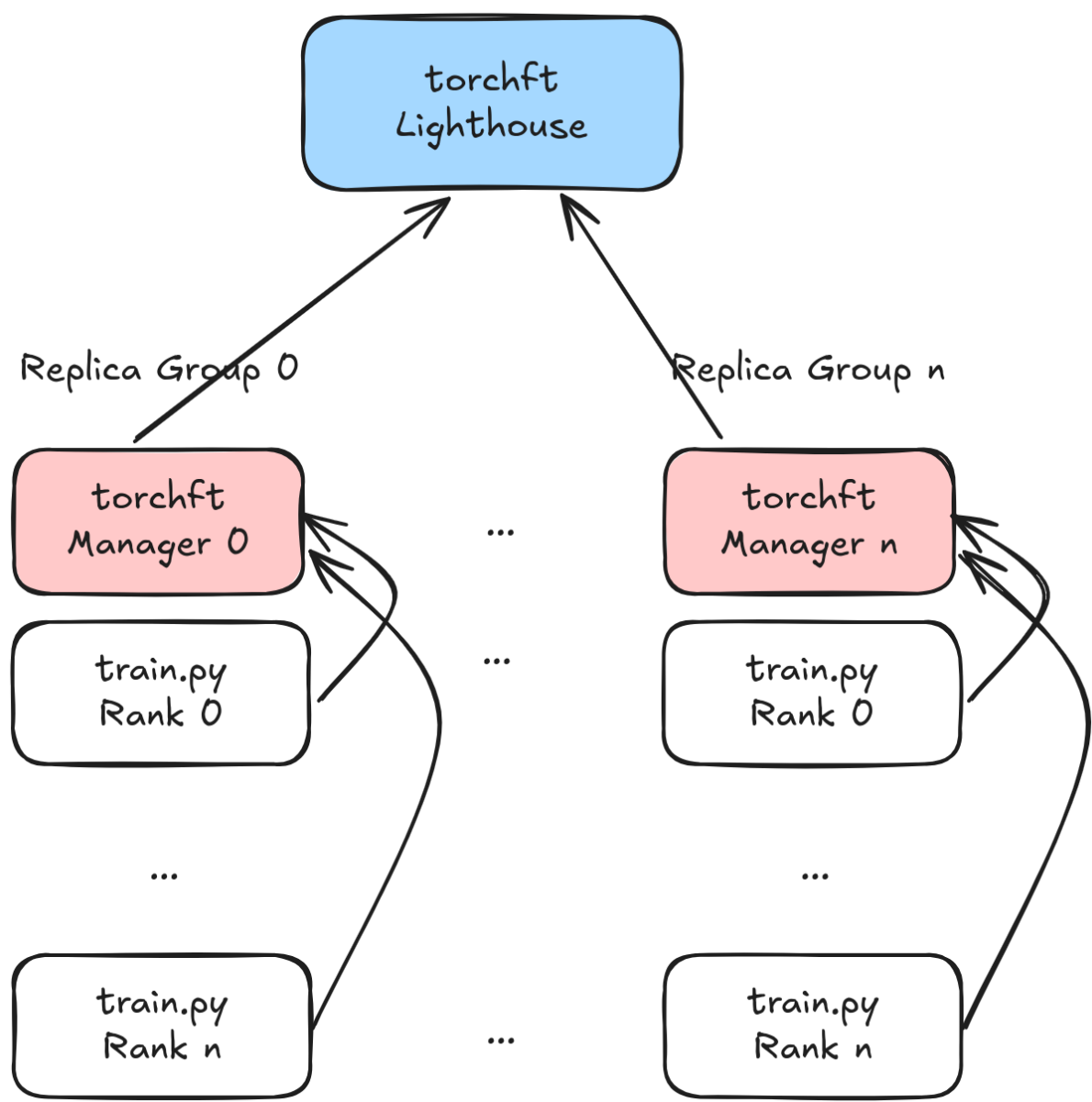
SageMaker HyperPod Infrastructure

- Automatic Hardware Failure Detection:** Continuous monitoring identifies GPU/node failures within minutes
- Node Replacement:** Faulty nodes automatically replaced without manual intervention
- Auto-Resume Functionality:** Training jobs resume from checkpoints after failures
- Health-Monitoring Agent:** Proactive infrastructure repair before failures impact training
- Integrated Observability:** Consolidated dashboard on GPU/Network/File System health, resource utilization, and task-level performance.



TorchFT Fault Tolerance

- Lighthouse Coordination Servers:** Managing replica groups with heartbeat-based health monitoring
- Per-Step Health Detection:** Identifies healthy workers at each training step
- Manager Component:** Orchestrates fault-tolerant gradient allreduce operations
- Quorum Consensus:** Maintains consistency across distributed workers
- Reconfigurable ProcessGroups:** Dynamic membership changes without stop-the-world operations



```
import torchft
from torchft.manager import Manager

# Initialize fault-tolerant manager
manager = Manager(
    min_replica_size=0.8,
    heartbeat_interval=10,
    quorum_timeout=30
)

# Fault-tolerant training loop
for step in range(num_steps):
    with manager.step():
        loss = model(batch)
        loss.backward()
    # Fault-tolerant allreduce
    manager.allreduce_gradients()
    optimizer.step()
```

Integration Workflow

```
apiVersion: apps/v1
kind: Deployment
metadata:
  name: torchft-lighthouse
spec:
  replicas: 1
  selector:
    matchLabels:
      app: torchft-lighthouse
  template:
    metadata:
      labels:
        app: torchft-lighthouse
    spec:
      nodeSelector:
        node.kubernetes.io/instance-type: "ml.g5.2xlarge"
        sagemaker.amazonaws.com/node-health-status: "Schedulable"
      containers:
        - name: lighthouse
          image: <ECR_REGISTRY>/torchft:latest
          command:
            - /bin/bash
            - "RUST_BACKTRACE=1 torchft_lighthouse --min_replicas 1 --quorum_tick_ms 100 \
              --join_timeout_ms 10000 --bind 0.0.0.0:29510"

ports:
  - containerPort: 29510
apiVersion: apps/v1
kind: StatefulSet
metadata:
  name: torchft-training
  annotations:
    sagemaker.amazonaws.com/enable-job-auto-resume: "true"
    sagemaker.amazonaws.com/job-max-retry-count: "5"

spec:
  serviceName: torchft-training
  replicas: 2
  selector:
    matchLabels:
      app: torchft-training
  template:
    metadata:
      labels:
        app: torchft-training
    spec:
      restartPolicy: Always
      nodeSelector:
        node.kubernetes.io/instance-type: "ml.p5en.48xlarge"
        sagemaker.amazonaws.com/node-health-status: "Schedulable"
      containers:
        - name: pytorch
          image: <ECR_REGISTRY>/torchft:latest
          resources:
            requests:
              nvidia.com/gpu: 8
              vpc.amazonaws.com/efa: 32
            limits:
              nvidia.com/gpu: 8
              vpc.amazonaws.com/efa: 32
          env:
            - name: TORCHFT_LIGHHOUSE
              value: "http://torchft-lighthouse:29510"
            - name: FI_PROVIDER
              value: "efa"
            - name: NCCL_DEBUG
              value: "INFO"
            - name: NCCL_SOCKET_IFNAME
              value: "^lo"

            - name: NUM_REPLICA_GROUPS
              value: "2"

            - name: PYTHONUNBUFFERED
              value: "1"
            command:
              - /bin/bash
              - -c
              - |
                REPLICA_GROUP_ID=${HOSTNAME##*-}
                export REPLICA_GROUP_ID
                MASTER_PORT=$((29600 + REPLICA_GROUP_ID))

                torchrun \
                  --master_port=$MASTER_PORT --nnodes=1 \
                  --nproc_per_node=8 train.py
```

Zero
Manual
Intervention

95%+
GPU Utilization

Weeks
Continuous
Training

Normal Operation (2 Nodes)

[INFO] [torchft::lighthouse] - Quorum status: Fast quorum found! [2/2 participants healthy][2 heartbeating][shrink_only=false]
[INFO] [torchft::lighthouse] - Detected quorum change, bumping quorum_id to 6

Inject XID Error to Trigger Node Failure

```
sudo sh -c "echo \"$(date '+%b %d %H:%M:%S') $(hostname) kernel: NVRM: Xid (PCI:0000:b9:00): 0, mock error\" >> /var/log/messages"
```

HyperPod detects faulty node and begins replacement workflow

[INFO] [HMA] Instance i-0XXXXXXX is unhealthy. HMA detected fault type NvidiaErrorTerminate
[INFO] [HMA] EC2 Instance i-0XXXXXXX successfully terminated

TorchFT detects change and updates quorum

[INFO] [torchft::lighthouse] - Quorum status: Valid quorum found [1/1 participants healthy]
[INFO] [torchft::lighthouse] - Detected quorum change, bumping quorum_id to 7

HyperPod automatically provisions replacement instance

[INFO] [HMA] Instance creation started successfully
[INFO] [HMA] EC2 Instance i-0XXXXXXX successfully provisioned

TorchFT detects change and updates quorum

[INFO] [torchft::lighthouse] - Quorum status: Fast quorum found! [2/2 participants healthy][2 heartbeating][shrink_only=false]
[torchft::lighthouse] - Detected quorum change, bumping quorum_id to 8

