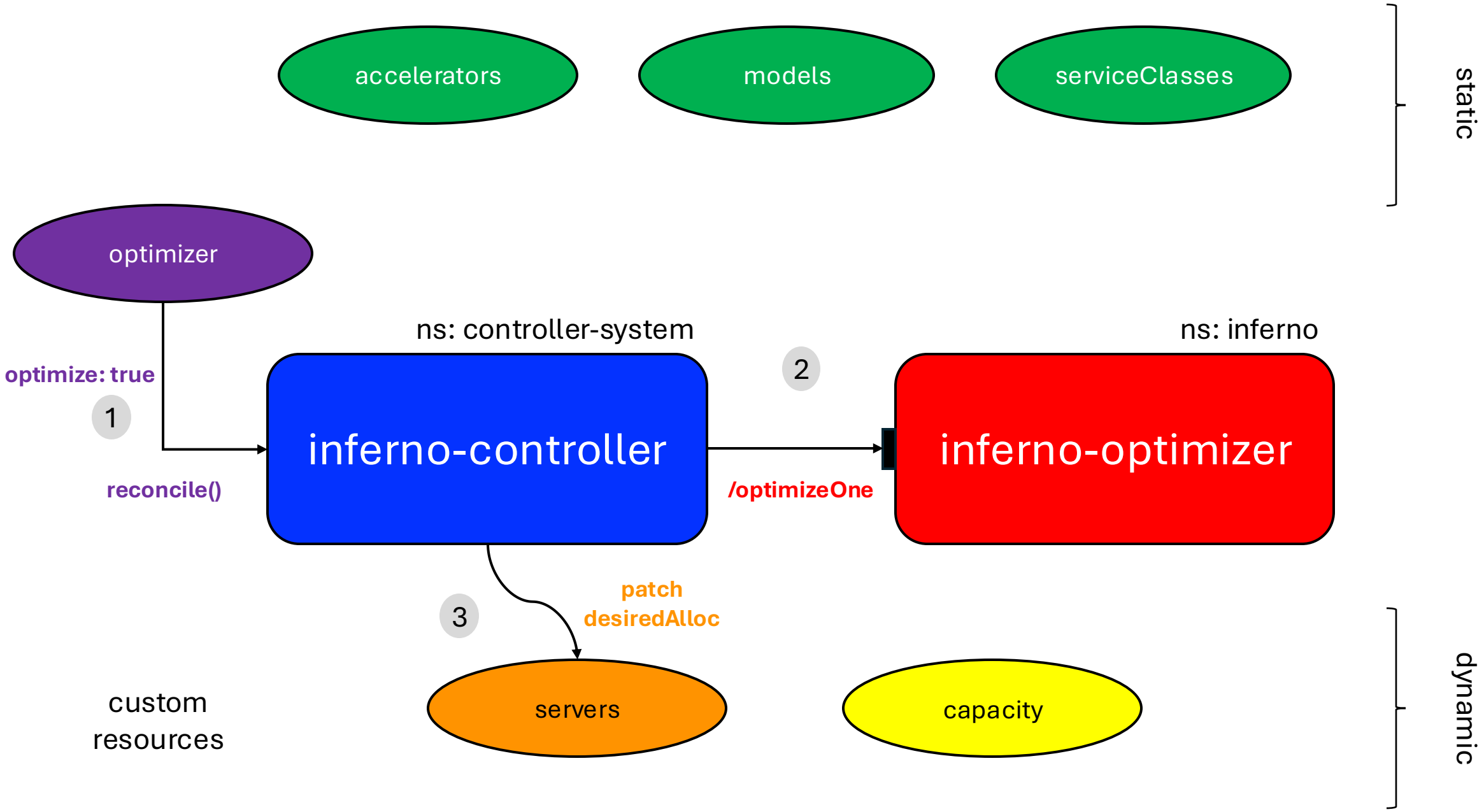
accelerators
models
serviceClasses
static
optimizer
ns: controller-system
ns: inferno
optimize: true
1
2
inferno-controller
inferno-optimizer
reconcile()
/optimizeOne
3
patch
desiredAlloc
servers
capacity
custom
resources
dynamic

```yaml
apiVersion: inferno.platform.ai/v1alpha1
kind: Accelerator
metadata:
  name: a100
spec:
  name: "A100"
  type: "A100"
  multiplicity: 1
  power:
    idle: 150
    full: 400
    midPower: 320
    midUtil: 0.6
  cost: 40.00
```

```yaml
apiVersion: inferno.platform.ai/v1alpha1
kind: Model
metadata:
  name: granite-13b
spec:
  name: "granite-13b"
  data:
    - "acc": "A100"
      "accCount": 1
      "alpha": 20.58
      "beta": 0.41
      "maxBatchSize": 32
      "atTokens": 512
    - "acc": "G2"
      "accCount": 1
      "alpha": 17.15
      "beta": 0.34
      "maxBatchSize": 38
      "atTokens": 512
```

```yaml
apiVersion: inferno.platform.ai/v1alpha1
kind: ServiceClass
metadata:
  name: premium
spec:
  name: "Premium"
  "priority": 1
  "data":
    - "model": "granite-13b"
      "slo-itl": 40
      "slo-ttw": 500
    - "model": "llama0-70b"
      "slo-itl": 80
      "slo-ttw": 500
```

```yaml
apiVersion: inferno.platform.ai/v1alpha1
kind: Server
metadata:
  name: premium-granite-13b
spec:
  name: "Premium-granite-13b"
  class: "Premium"
  model: "granite-13b"
  currentAlloc:
    accelerator: "A100"
    numReplicas: 4
    maxBatch: 16
    cost: 1600
    itlAverage: 25.2
    waitAverage: 726.5
    load:
      arrivalRate: 40
      avgLength: 1024
      arrivalCOV: 1
      serviceCOV: 1
  desiredAlloc:
    accelerator: ""
    numReplicas: 0
    load:
      arrivalRate: 0
      avgLength: 0
```

```yaml
apiVersion: inferno.platform.ai/v1alpha1
kind: Optimizer
metadata:
  name: inferno
spec:
  optimize: false
  data:
    optimizer:
      unlimited: true
      heterogeneous: false
      milpSolver: false
      useCplex: false
```

```yaml
apiVersion: inferno.platform.ai/v1alpha1
kind: Capacity
metadata:
  name: capacity
spec:
  count:
    - type: "A100"
      count: 4
    - type: "G2"
      count: 8
```