Capacity-aware Dynamic Volume Provisioning For LVM-based Local Storage

Dec. 7th, 2022

Cybozu, Inc.
Satoru Takeuchi

Agenda

- Motivation
- What is TopoLVM
- How TopoLVM works
- What’s next

Agenda

- **Motivation**
- What is TopoLVM
- How TopoLVM works
- What's next

About Cybozu

- A leading cloud service provider in Japan
- Providing software that supports teamwork

Cybozu's Kubernetes cluster

- On-premises K8s cluster
- Storage
  - Distributed Block&Object Storage
    - => Rook/Ceph
  - **Local fast(NVMe SSD) storage**
    - => ???

Requirements for local storage

- Users can create arbitrary sized volumes
  - Fixed size disks/partitions are inconvenient
- Volumes should be spread over nodes based on free storage capacity
  - Use storage capacity for each node evenly

What was the best storage driver?

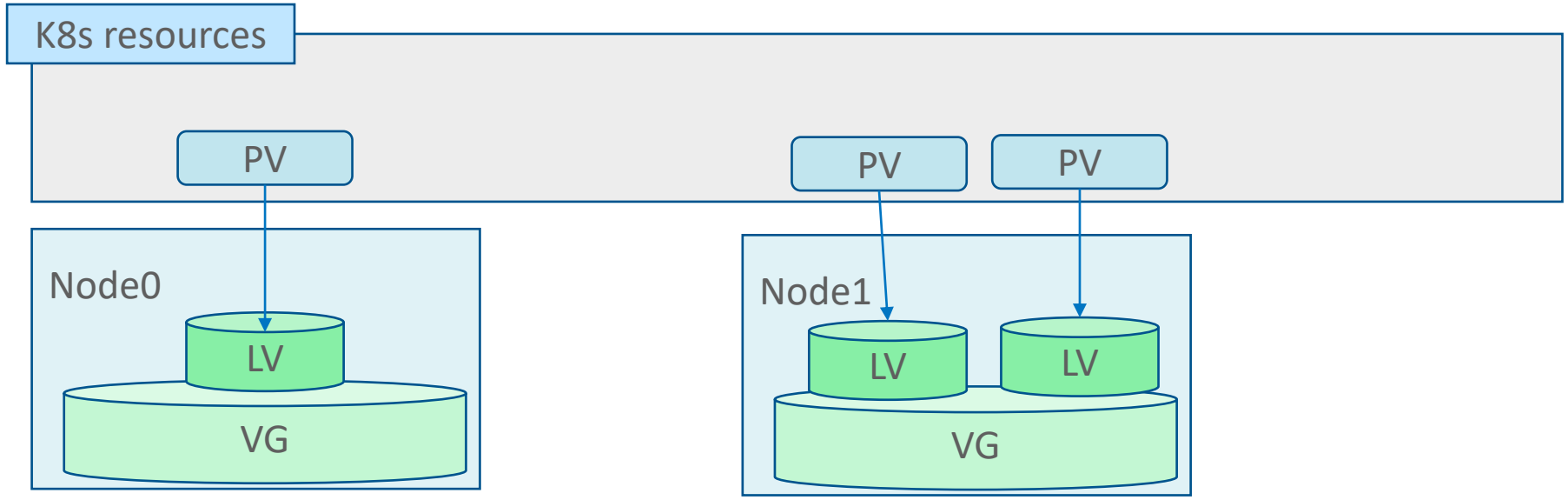
- There was no CSI driver that met all our requirements
- Decided to create a new CSI driver, TopoLVM

Agenda

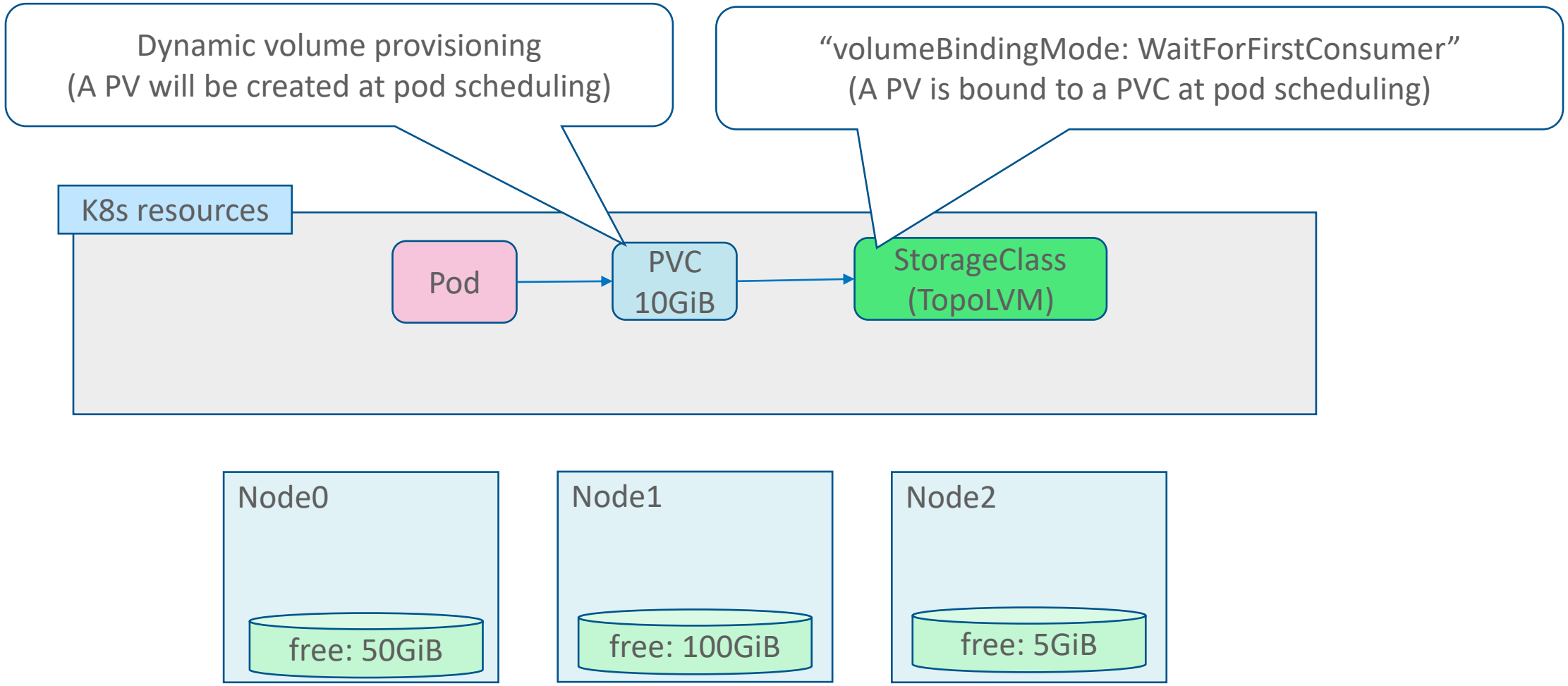
- Motivation
- **What is TopoLVM**
- How TopoLVM works
- What’s next

Arbitrary volume size

- TopoLVM deals with LVM VGs prepared on nodes
- TopoLVM creates an LVM LV for each PV resource

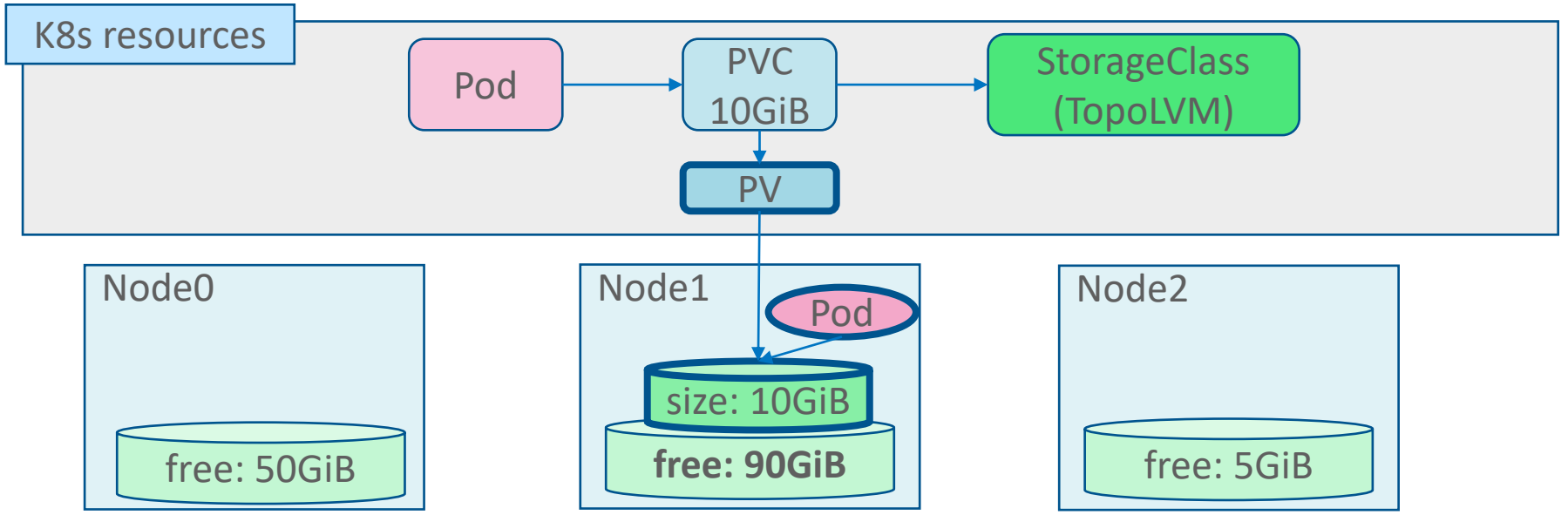


Pod scheduling and volume provisioning(1/2)



Pod scheduling and volume provisioning(2/2)

- The pod is scheduled to the node having the largest free VG space as possible (in this case, node1)
- The volume is provisioned on the same node (node1)



Other features

- ext4, XFS, Btrfs, and Raw Block Volume
- Generic ephemeral volume
- Volume expansion
- Thin volume
  - With thin snapshot and thin clone

Community

- There are many non-Cybozu users/developers
- Some companies use TopoLVM in their products

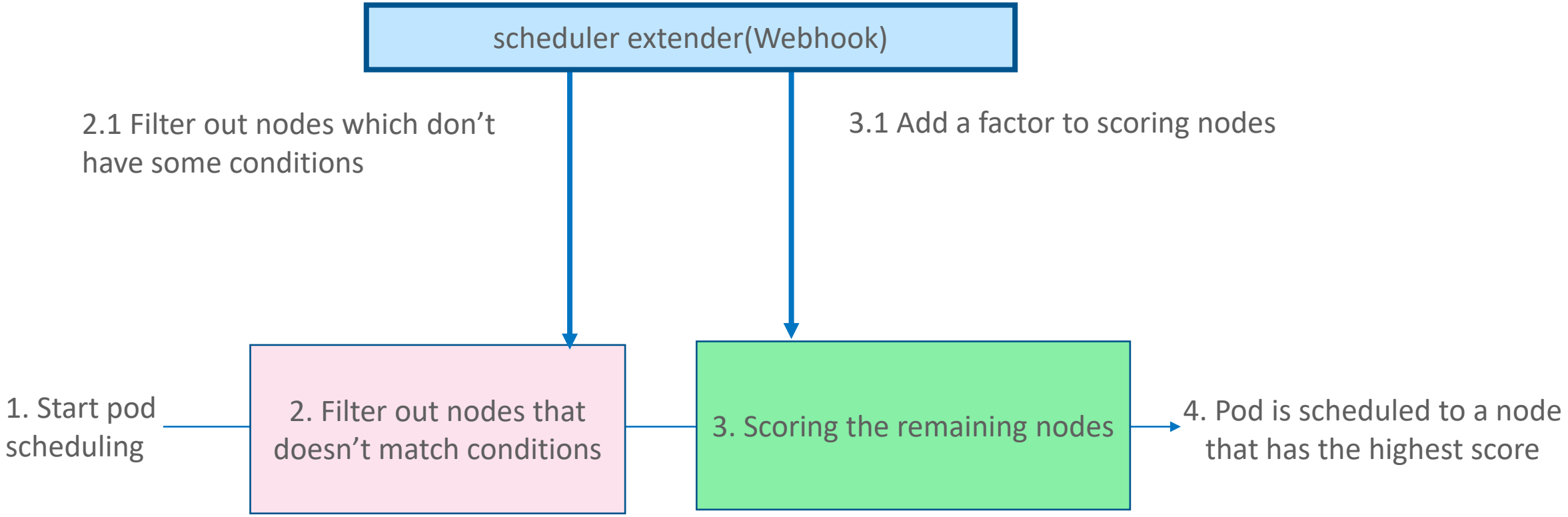
Agenda

- Motivation
- What is TopoLVM
- **How TopoLVM works**
- What’s next

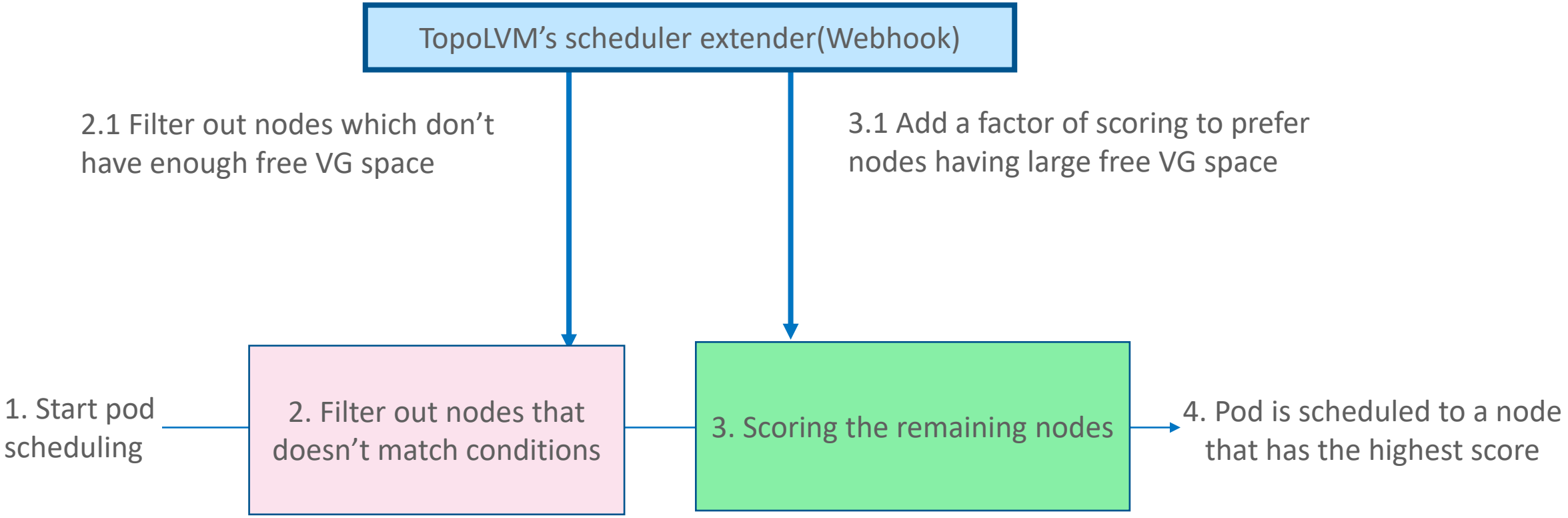
Challenges

- Schedule a pod to the node having as large free VG space as possible
  - => Scheduler extender
- Provision the volume on the same node
  - => CSI Topology

Scheduler extender



TopoLVM's scheduler extender



The parameters of the TopoLVM's scheduler extender

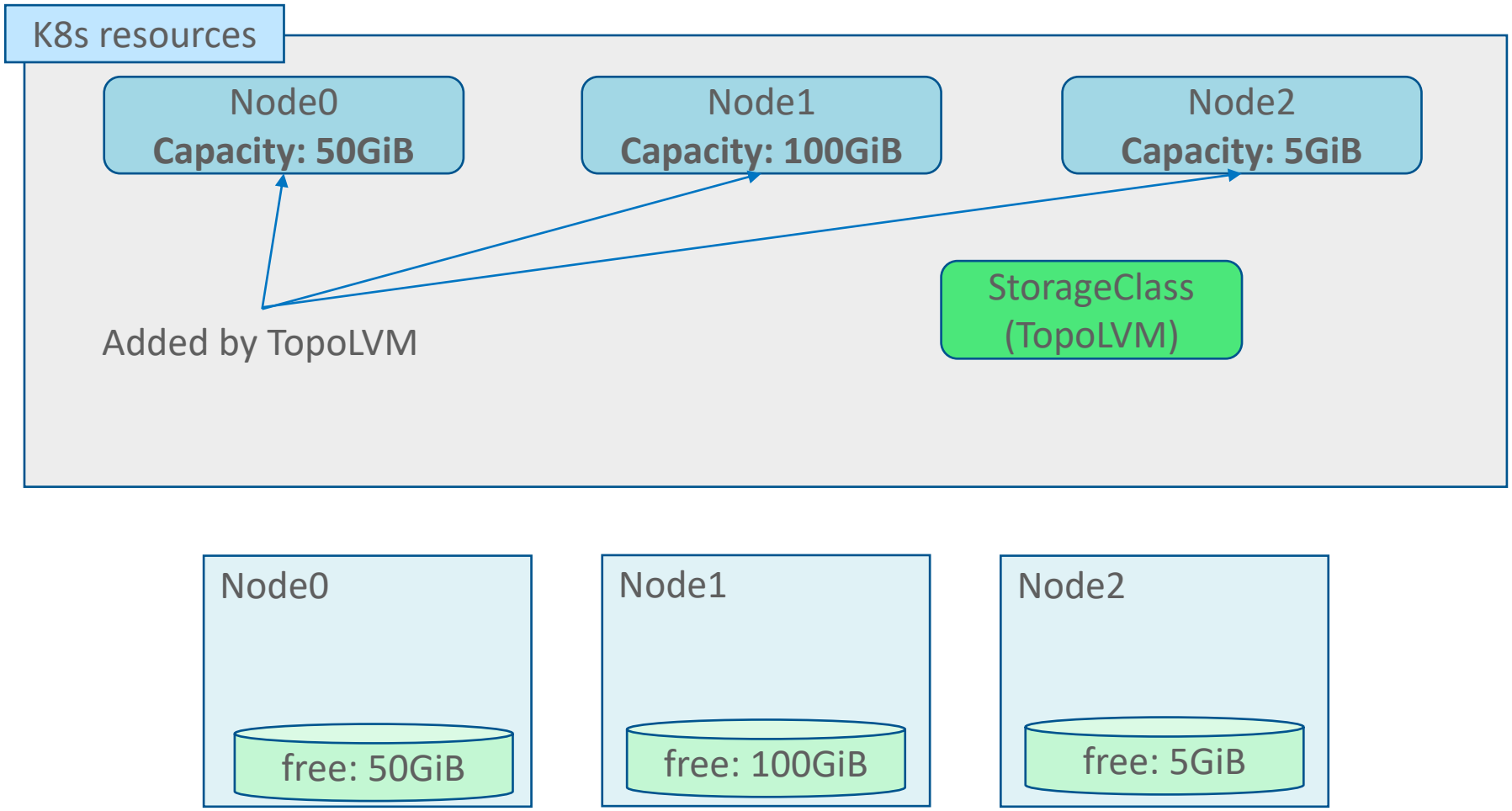
- TopoLVM's scheduler extender requires two kinds of parameters
  - Free VG space for each node(*1)
  - Total requested TopoLVM volume size for each Pod
- TopoLVM manages annotations for these parameters in node and pod resources

*1 K8s's StorageCapacityTracking feature can also be used only for filtering

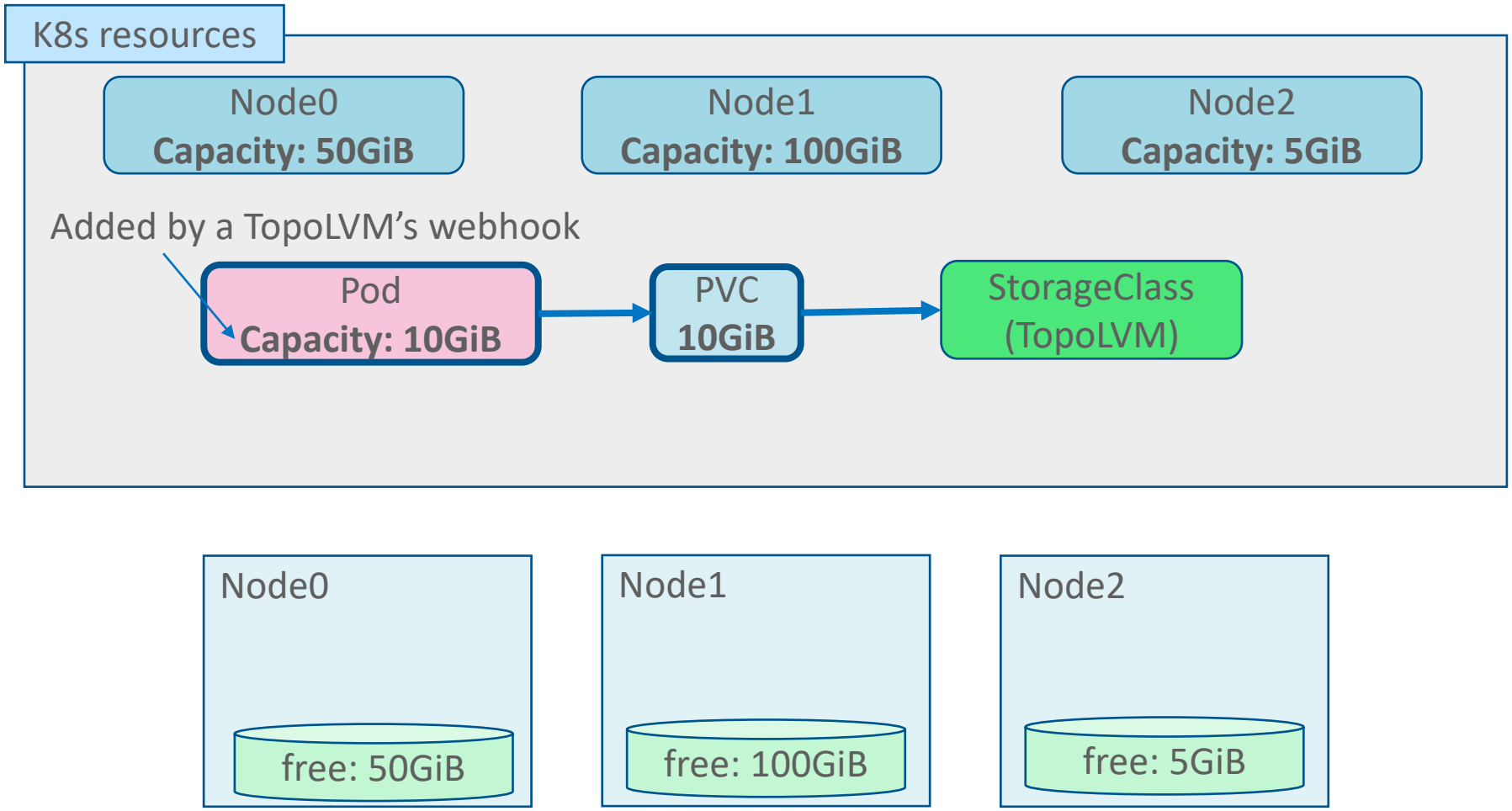
CSI topology

- A feature of Kubernetes
  - https://kubernetes-csi.github.io/docs/topology.html
- Schedule a pod to one of the nodes where its volumes are available
  - Used for zone local storage, node local storage, and so on
- TopoLVM create a volume on the same node as the corresponding pods.

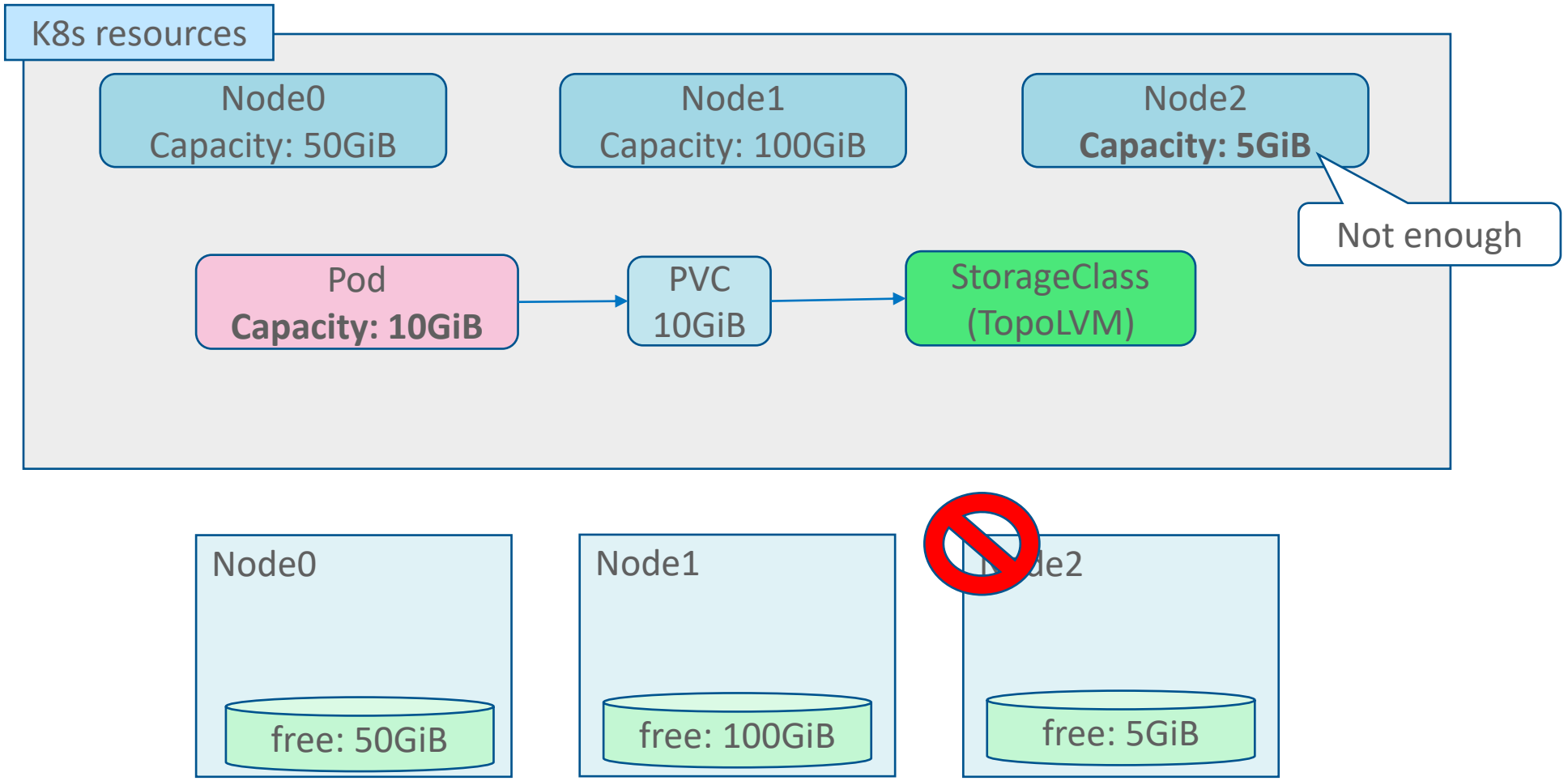
Example


K8s resources
Node0
Capacity: 50GiB
Node1
Capacity: 100GiB
Node2
Capacity: 5GiB
Added by TopoLVM
StorageClass
(TopoLVM)
Node0
free: 50GiB
Node1
free: 100GiB
Node2
free: 5GiB

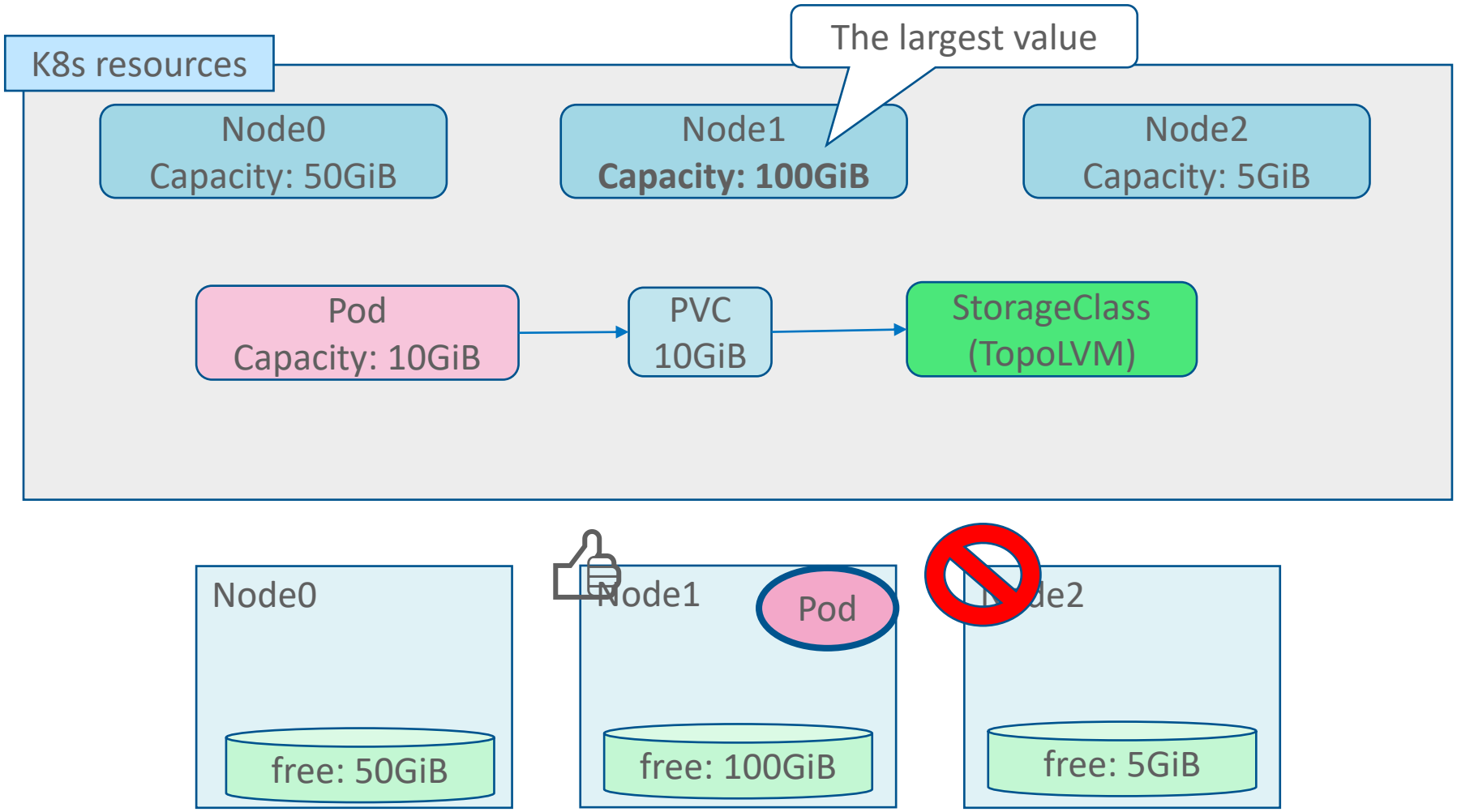
Create both Pod and PVC resources



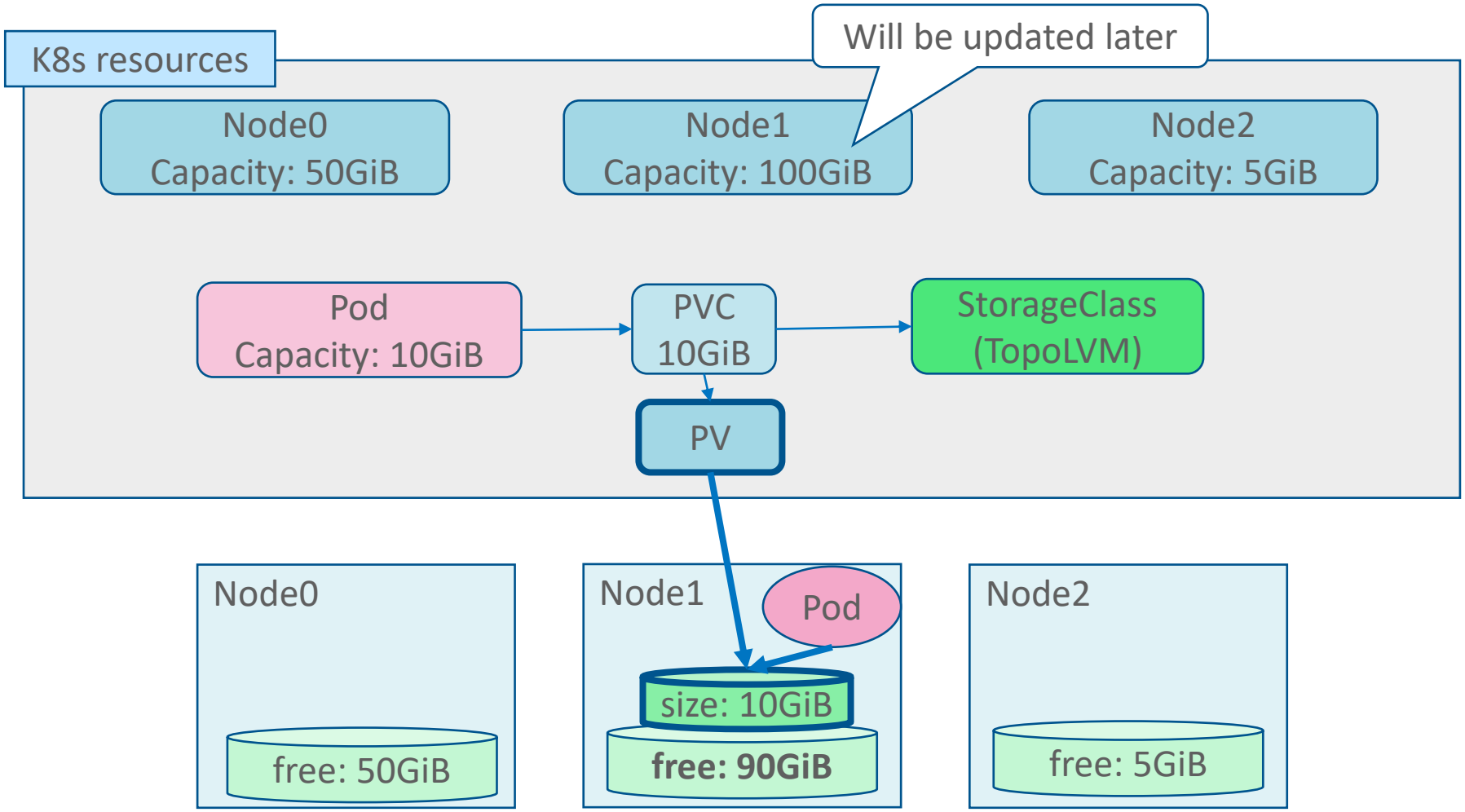
Scheduler extender: Filtering



Scheduler extender: Scoring



Provision and binding the volume



Agenda

- Motivation
- What is TopoLVM
- How TopoLVM works
- **What’s next**

Next plans

- Implement the K8s-official capacity-aware pod scheduling
  - Setting up a scheduler extender is a bit difficult
  - We're preparing a KEP
- Donate TopoLVM project to CNCF

Conclusion

- TopoLVM is an LVM-based CSI driver
- Volumes and the corresponding pods are evenly spread for each node
  - By scheduler extender and CSI topology
- We welcome new users and contributions

That’s all, thank you!

- Project page
  - https://github.com/topolvm/topolvm
- A blog post about TopoLVM
  - https://blog.kintone.io/entry/topolvm