



## HIGHLIGHTS

- ▶ Efficiently coordinates a set of Horizontal Pod Autoscalers (HPAs)
- ▶ Proactively addresses infrastructure latency
- ▶ Employs forward-thinking optimization strategies, rather than short-term, greedy approaches.
- ▶ Features event-based scaling for dynamic resource allocation.
- ▶ Manages load balancing across multiple clusters for optimal performance.
- ▶ Enables autonomous scaling operations, reducing manual intervention.
- ▶ Offers predictive remediation to anticipate and resolve issues before they escalate.
- ▶ Metrics Supported:
  - ▶ Application Metrics:
    - ▶ Requests Per Second (RPS)
    - ▶ Latency
    - ▶ Faults
  - ▶ Infrastructure Metrics:
    - ▶ Pod Count
    - ▶ CPU Utilization
    - ▶ Memory Usage
  - ▶ Custom Metrics:
    - ▶ Temperatures
- ▶ Data Sources:
  - ▶ Prometheus
  - ▶ Datadog
  - ▶ New Relic
  - ▶ Dynatrace

## ▶ Application scaling comparison

Comparing traditional methods with Generative AI [RL] powered Smart Scaler

|                                      | Vanilla HPA                      | Keda                | Smart Scaler                 |
|--------------------------------------|----------------------------------|---------------------|------------------------------|
| <b>Predictive Autoscaling</b>        | NA                               | NA                  | Yes with Application Metrics |
| <b>Calendar Aware Scaling</b>        | None                             | None                | First to market              |
| <b>Service Graph Aware Scaling</b>   | None                             | None                | Service Graph Aware          |
| <b>DevOps Tuning</b>                 | manual + experience              | manual + experience | Gen AI                       |
| <b>Accuracy</b>                      | Best Guess                       | Better              | guaranteed                   |
| <b>DevOps Resource Effectiveness</b> | Poor                             | Better              | Fantastic                    |
| <b>Effort</b>                        | Days, Nights, Weekends, Holidays | Weeks               | 30 Mins... tops              |
| <b>DevOps Burnout</b>                | Lots                             | Lots                | Proactive                    |
| <b>Cost</b>                          | overprovisioning                 | overprovisioning    | 15-50% Savings               |
| <b>Resource Spin Up Time</b>         | reactive                         | reactive            | Predictive                   |
| <b>Low error rate</b>                | Hard to achieve                  | Hard to achieve     | .001% Errors                 |