November 2020

Powering AI and Analytics with HPC

Jeff Reser, SUSE Solutions

SUSE

Contact.

Jeff Reser

Head of Portfolio Marketing

Jeff.Reser@suse.com
+1 919.500.1733

Alessandro Festa

Sr. Product Manager, AI/ML

Alessandro.Festa@suse.com
+39 335 75 54 151

SUSE

Agenda.

- Challenges and Strategy
- Use Cases and Outcomes
- Addressing the Challenges
- Learning More

AI/ML Challenges

Assertions.

AI/ML analytics turns data into actionable insights

Businesses will increasingly leverage machine learning

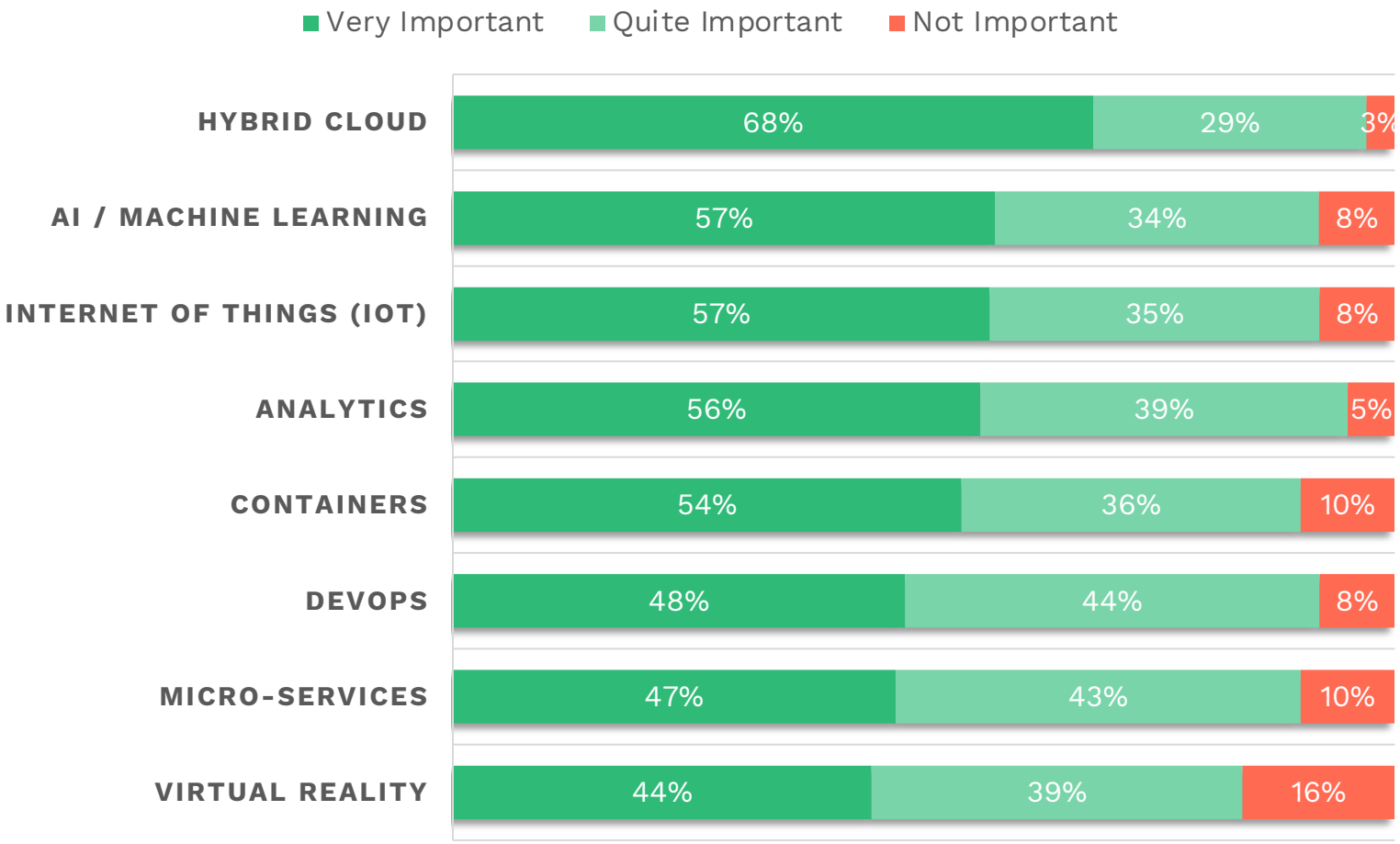
AI will morph into ‘Practical AI’ and become more useful in everyday life

Practical AI strengthens the trust people have in its inferences

Enabling Business Transformation

- Momentum High-Performance Computing, Edge Computing, AI and analytics reflects an appetite for innovation
- Hybrid cloud, Software-Defined Infrastructure and Container Management enable that journey

IMPORTANCE OF ENHANCING SKILLS AND EXPERIENCE BY AREA[1]



[1] SUSE White Paper "How Today's IT Leaders are Daring to be Different", May 2020

Considerations.

Strategy: AI success needs a clear, concise and adaptable strategy

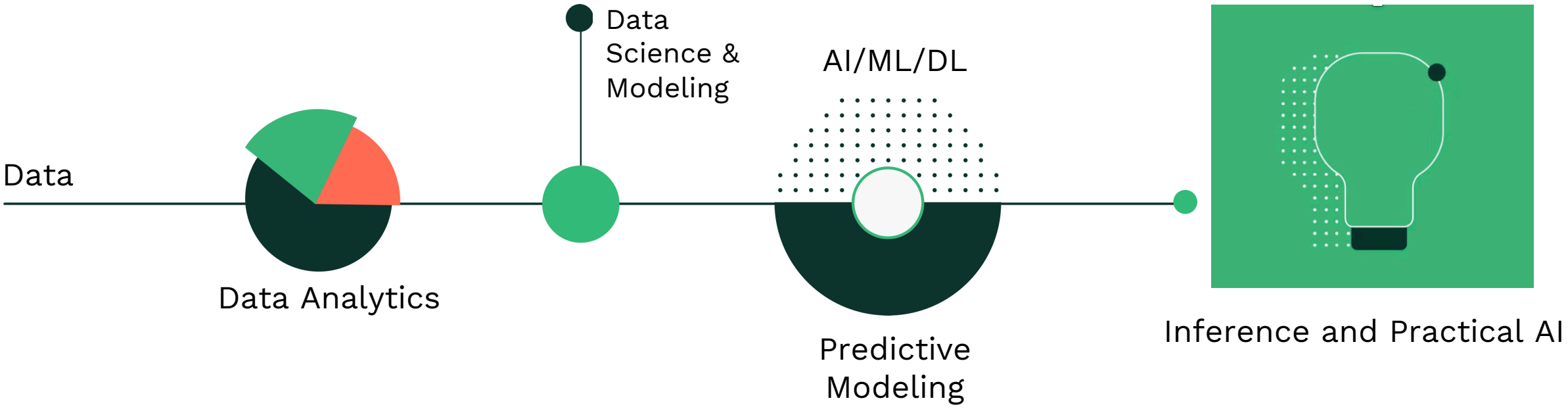
People: Right skills to leverage AI technology to transform the business

Process: Improve efficiencies in business and IT operations

Technology: Analyze large volumes of data using ML to unlock insights

Strategy

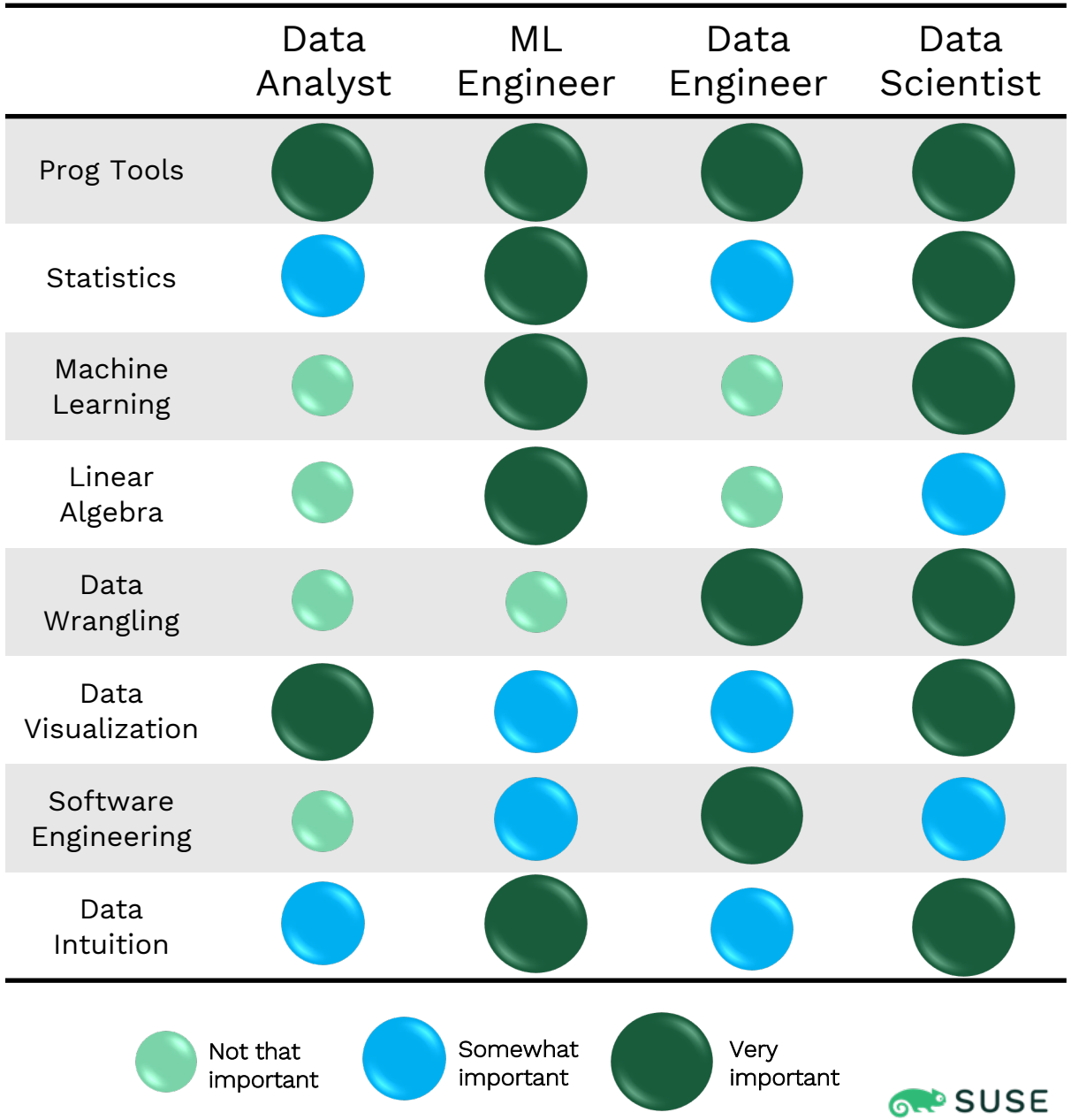
AI success needs a formalized investment plan



People

Skills to leverage AI technology to transform the business

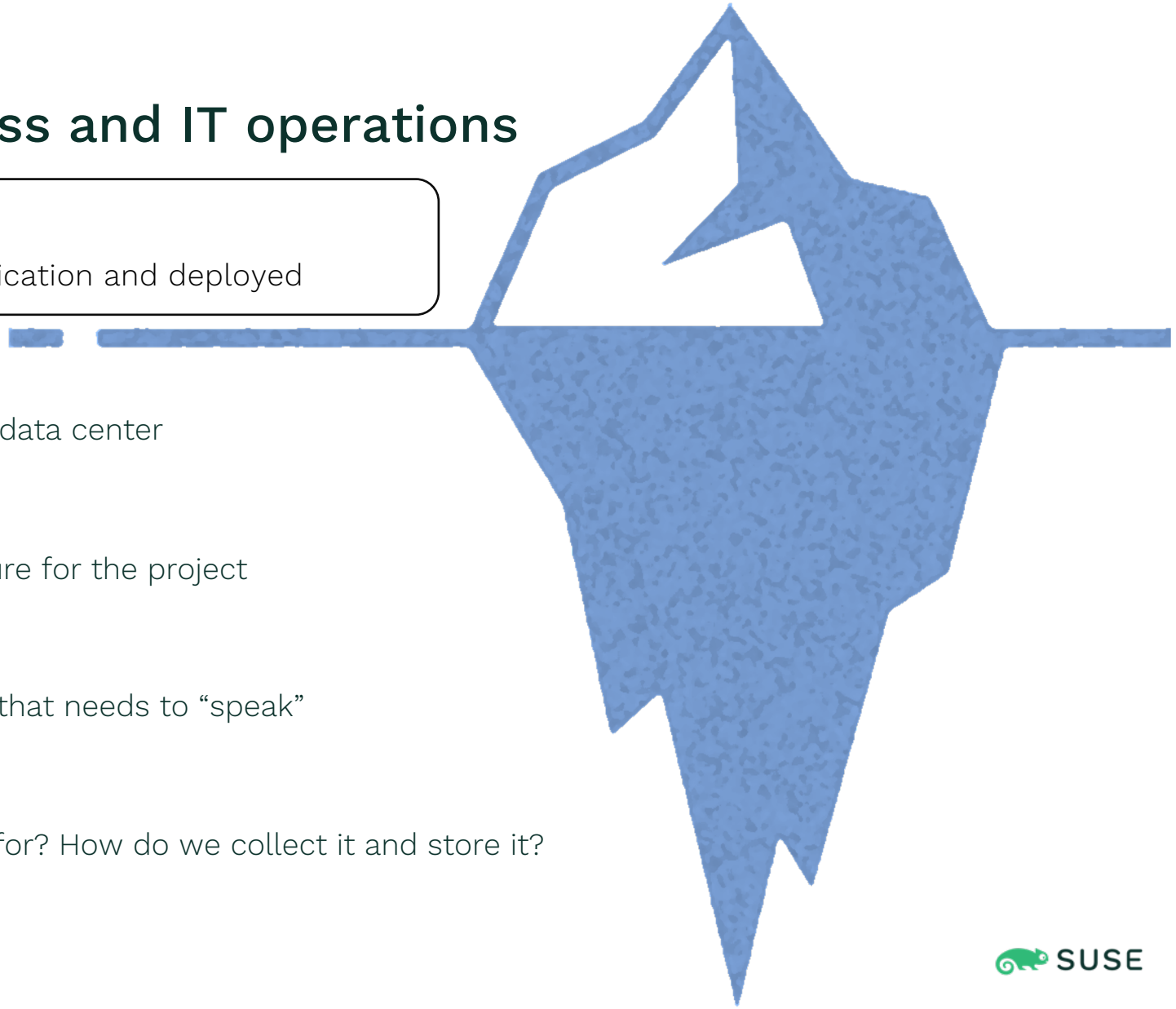
- Data science skills and competencies[1] are valuable for getting the most out of your data through AI/ML
- Skills needed to develop, implement and operate AI systems
- Besides internal skills, external consultants/suppliers with credible knowledge or experience are needed



[1] https://blog.udacity.com/2014/11/data-science-job-skills.html

Process

Improve efficiencies in business and IT operations



Production Model

Model is ready, injected in your application and deployed

Pipeline Execution

Model is running in a pipeline in your data center

Choose the right infrastructure

It's about having the right infrastructure for the project

Data Set Preparation

It's not just simply data, but it's data that needs to "speak"

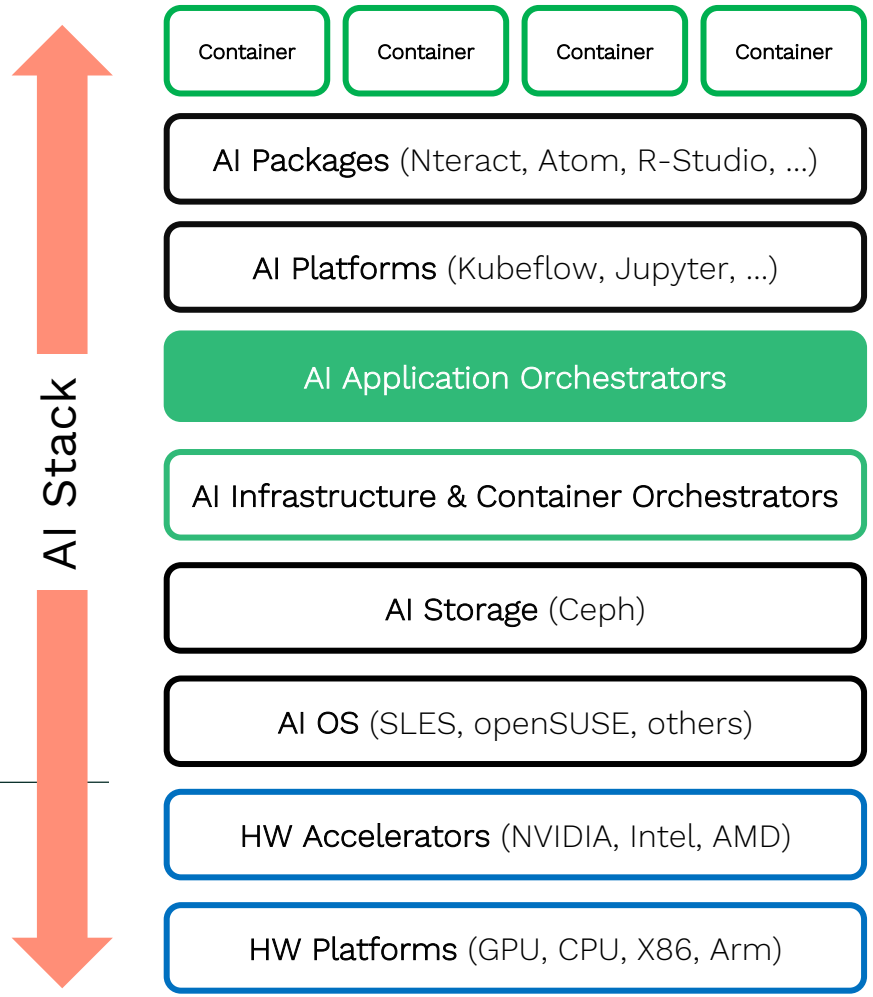
Data Gathering

What do we need? What do we look for? How do we collect it and store it?

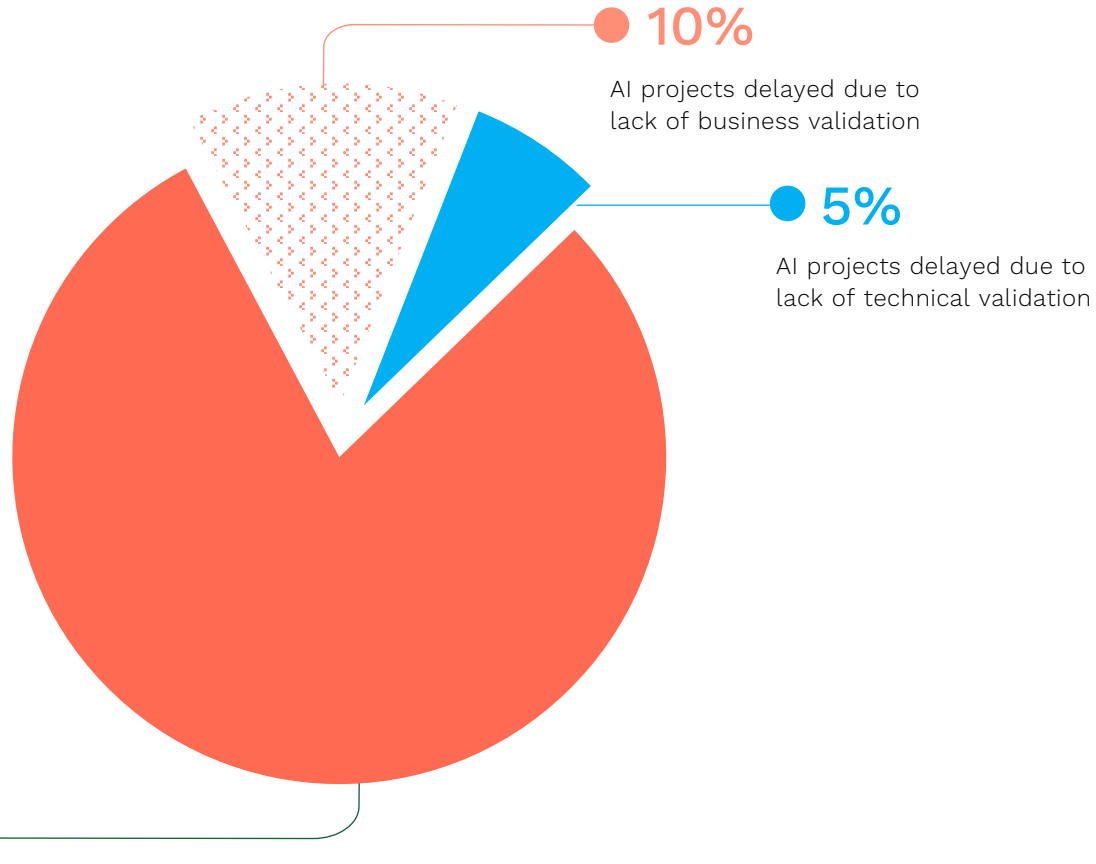
Technology

Analyze large volumes of data using ML to unlock insights

- End-to-end experience that reduces time to production
- Simplifies steps to move AI projects from prototype to test/validation to release
- Provides guidance on which technologies and products to use within each layer of the stack
- Represents an optimized infrastructure that perfectly fits the customers hardware architecture



The AI Project Dilemma



"Increasingly, organizations are looking for not just the hardware but a complete AI ***infrastructure stack*** *that combines server hardware, hardware abstraction layers, orchestration layers, AI development layers and data science layers that seamlessly operate together."*

- IDC "Worldwide AI Server Forecast, 2020-2024: COVID-19 Disrupts the Momentum", Peter Rutten

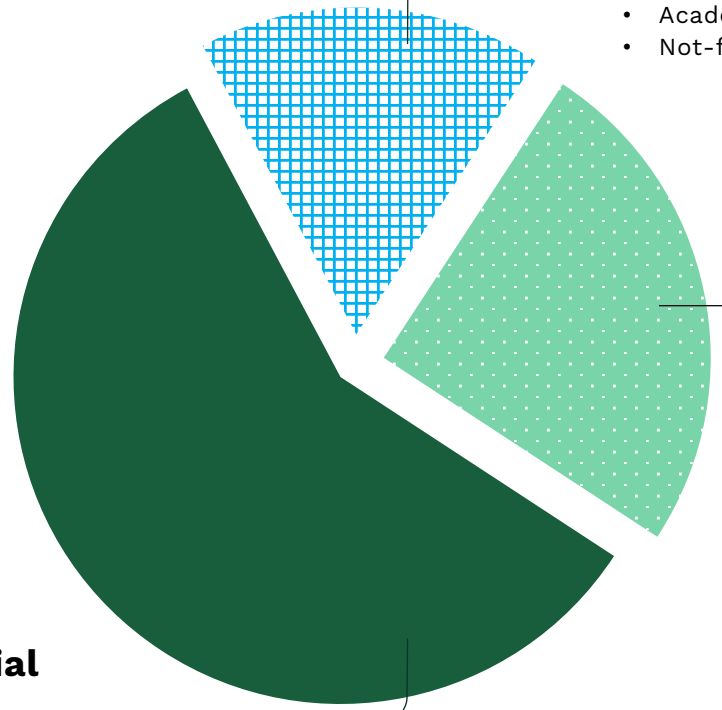
"Launching pilots is deceptively easy but deploying them into production is notoriously ***challenging ...*** *Although the potential for success is enormous, delivering business impact from AI initiatives takes much longer than anticipated ..."*

- Chirag Dekate, Gartner – "Gartner Predicts the Future of AI Technologies"

HPC Infrastructure Is Being Driven Into Enterprise Markets[1]

Competitive forces are driving companies to aim more complex questions at their data structures and push business operations closer to real time.

HPC moving in-house for scalability, ultrafast data movement and very large memory systems.



Academic

20%

- Academic
- Not-for-profit

Government

25%

- National agency
- National security
- National research lab
- State or local government

Commercial

58%

- Financial services
- Large product manufacturing
- Bio-sciences
- Energy
- Consumer product manufacturing
- Retail
- Chemical
- Media & Entertainment
- Electronics
- Transportation
- Other commercial

[1] *Intersect360 Research, November 2019*

SUSE

HPC Market Factoids

- HPC **ROI is very high** - $458 (on average) revenue per dollar; $58 average profit (or cost savings) per dollar invested in HPC[1]
- Worldwide HPC revenue expected to reach over **$19.95 billion** by 2023[1]
- Big data combined with HPC creating new solutions, adding many new users/buyers to the HPC space (**AI/ML/DL and HPDA** hot)
- **SUSE runs on 21 of the top 50** supercomputers (7 RH, 9 CentOS)[2]
- SUSE dominates top 100, CentOS gains share in "smaller" supercomputers[2]
- Commercial OS Share in Top 500 (represents 100 supercomputers in the list): **SUSE 53%**, RH 24%, bullx 17%, Ubuntu 6%[2]

[1] Hyperion Research, November 2019
[2] Top500 Supercomputer Report, November 2019

Hyperion Research Predictions[1]

- High growth rate of the HPC market continues
- HPC products being driven into broader enterprise market
- The exascale race will drive new technologies
- Many new processors and accelerators are on the way
- Storage systems will increasingly become more critical
- Cloud computing for HPC workloads will grow faster
- Artificial Intelligence will grow faster than everything else

[1] Hyperion Research, November 2019

Use Cases and Outcomes

Consumer Goods – Appliance Design

Outcomes using AI/ML

- **Manufacturing:** saves time/money, improves customer sat
- **Logistics:** saves money, fewer returns
- **Advertising:** creates personalized experiences

Energy & Utilities – Sustainable Energy

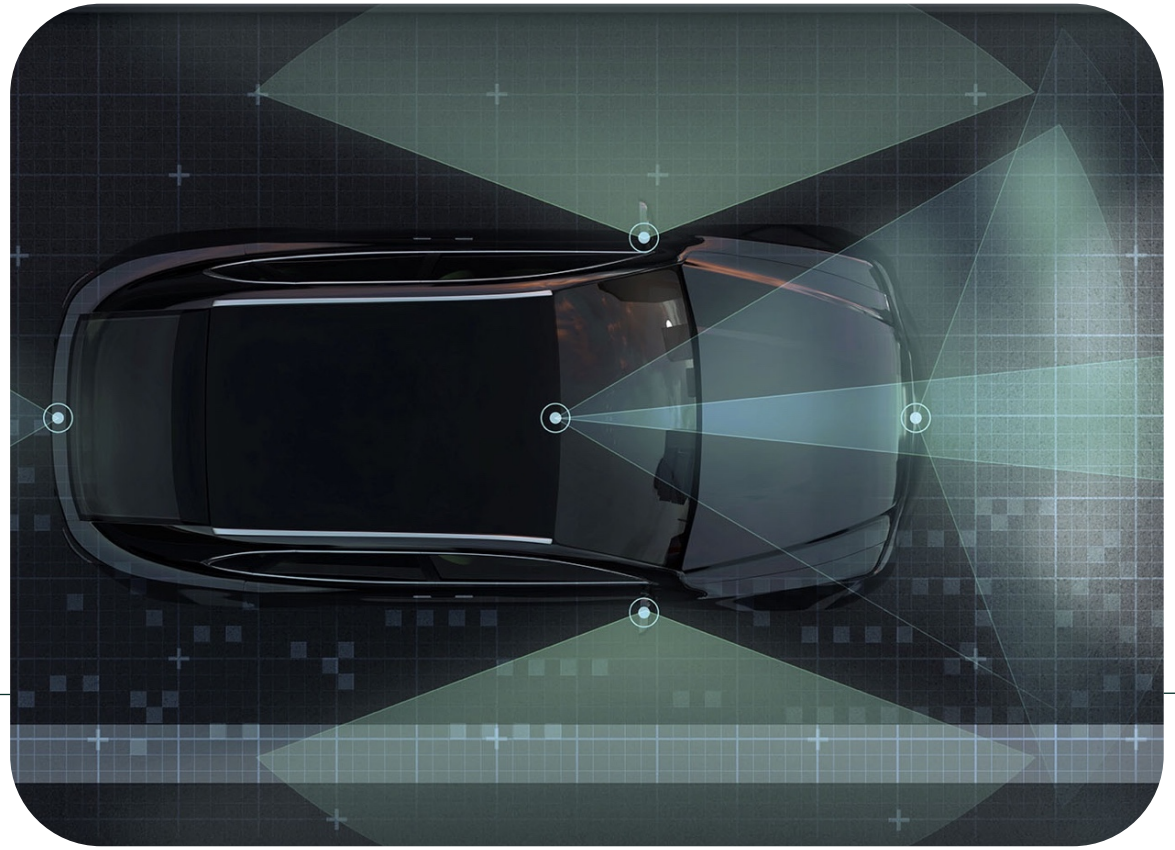
Outcomes using AI/ML

- **Modeling:** Reduces cost and risk
- **Insights:** Limits environment impacts; optimizes supply/demand; enables proactive maintenance
- **Efficiency:** Enables smart allocation of energy resources

Automotive – Design and Manufacturing

Outcomes using AI/ML

- **Design:** Enables effective design simulations
- **Connected vehicles:** Powers advanced safety features; cloud services for available data; driver monitoring
- **Manufacturing:** Robots drive optimization



Manufacturing – Materials Science

Outcomes using AI/ML

- **Discovery:** Discovers materials faster; mine databases for “recipes”
- **Analysis:** Predicts right compound combinations
- **Modeling:** Helps refine materials for optimum performance

Pharmaceuticals – Drug Research

Outcomes using AI/ML

- **Experimentation:** Predicts treatment results accurately
- **Discovery:** Improves drug design and discovery
- **Treatment:** Enables better disease management; enables precision medicine

SUSE AI Orchestrator

What is it?

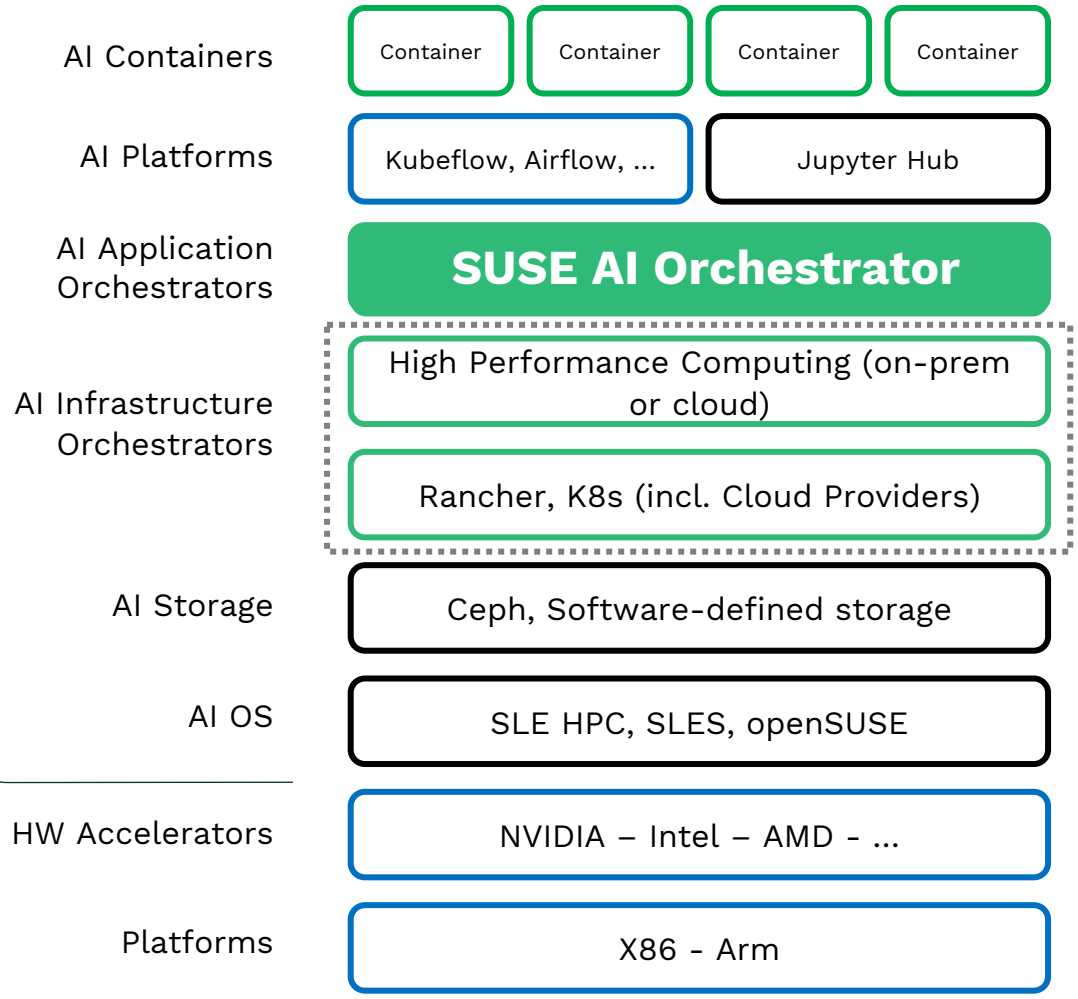
- A cloud-native tool that translates a data model into the execution steps of an AI platform pipeline or workflow in an automated way

How does it work?

- When the data scientist submits an execution request against the unmodified data model, the tool discovers and applies all the runtime options required for the chosen AI platform

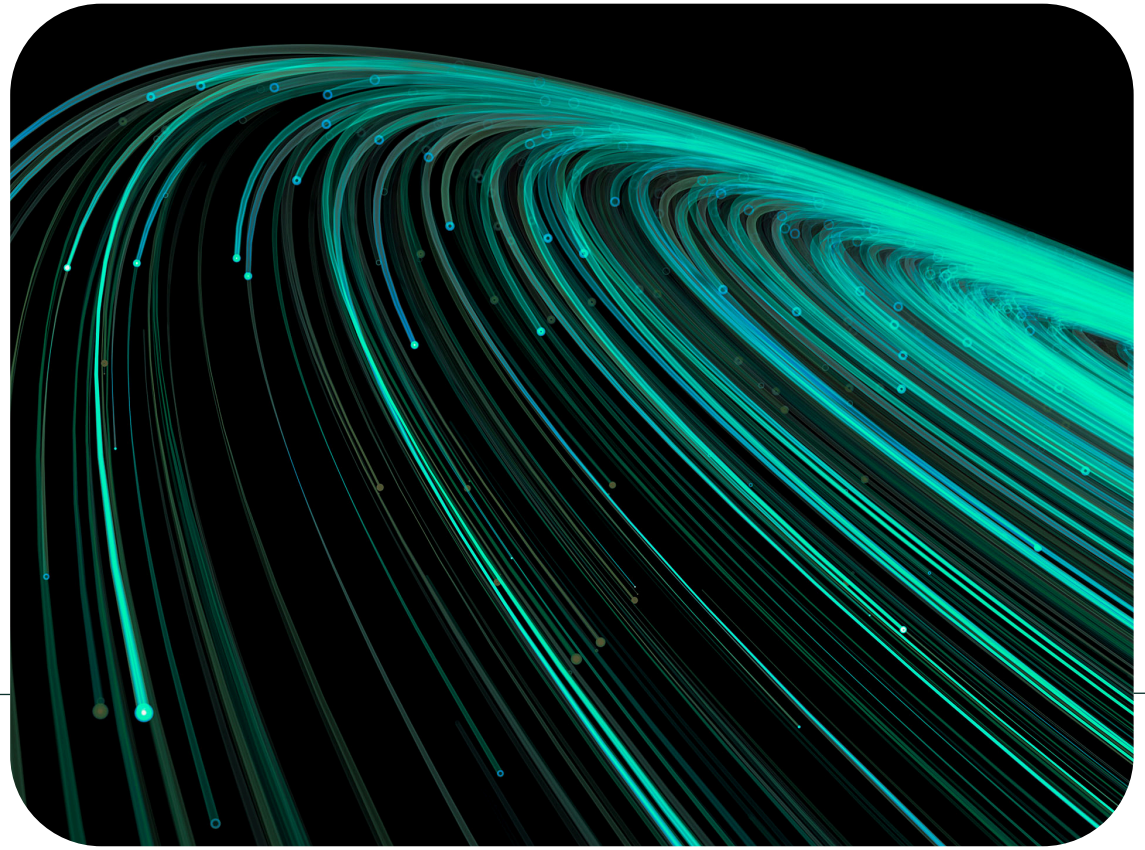
Why use it?

- **Automates** pipeline or workflow across AI platforms
- Fosters **collaboration** between data scientists and AI operators
- **Monitor or deploy** an entire AI platform on-premise or in the cloud



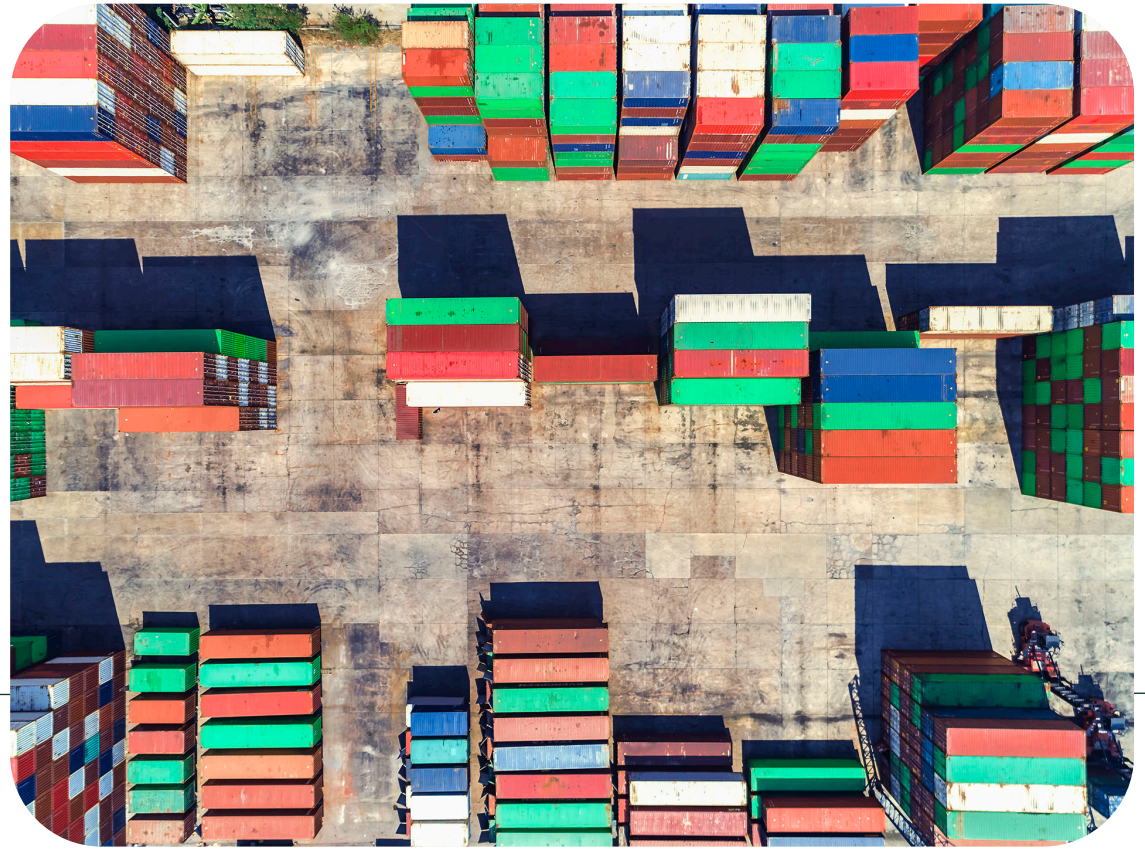
SUSE Linux Enterprise High Performance Computing

- Popular HPC tools and libraries bundled with SLE HPC
- All packages supported by SUSE
- Available for x86-64 and Arm64
- Flexible release schedule
- SLE 12 and SLE HPC 15



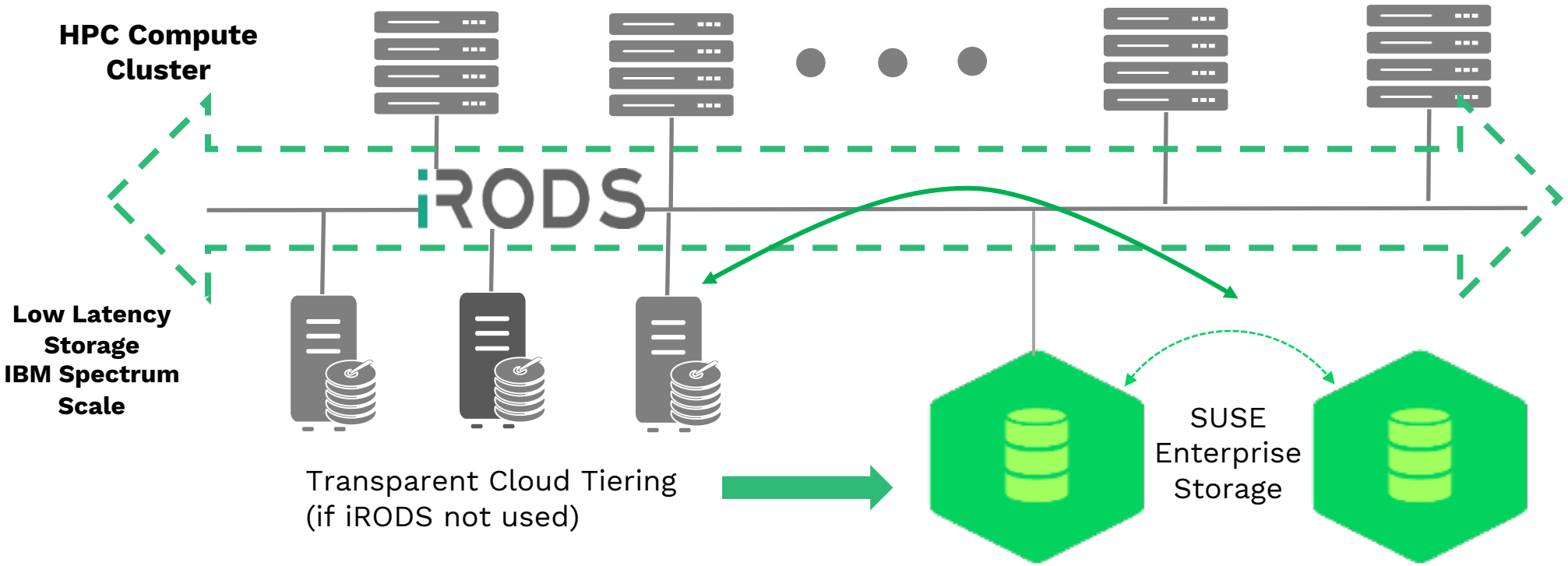
SUSE Enterprise Storage

- SUSE Enterprise Storage
- Ceph-based, software-defined
- Backup/archival HPC storage
- IO500 benchmark-ranked #17 (November 2019), showcasing HPC storage performance
- Easy to manage with the Ceph Dashboard



Tiered HPC Storage

Common Use Case – Tier 2 Storage/Active Archive

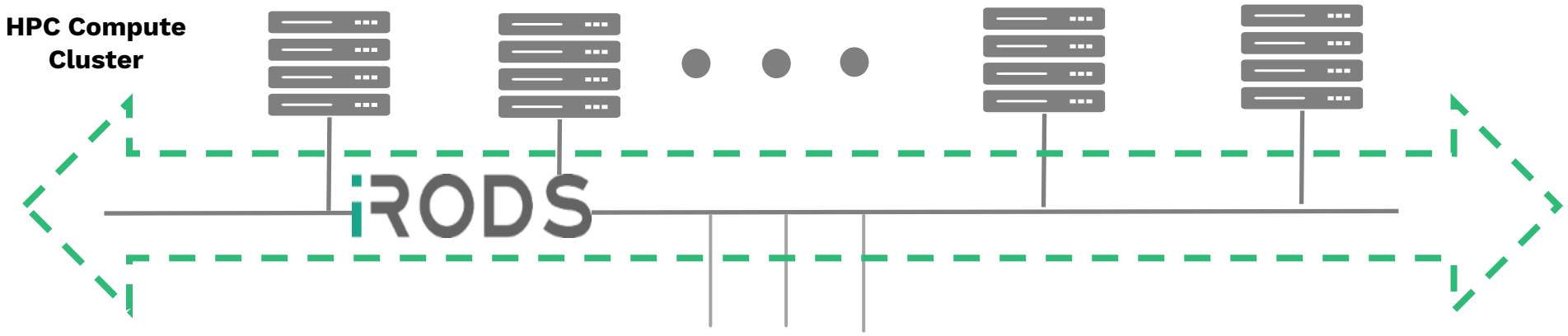


iRODS storage tiering

- Data migration based on pre-specified rules from primary to secondary storage
- Data landing zone provides fast tier of storage for incoming stream of data
- SUSE Enterprise Storage for longer term storage

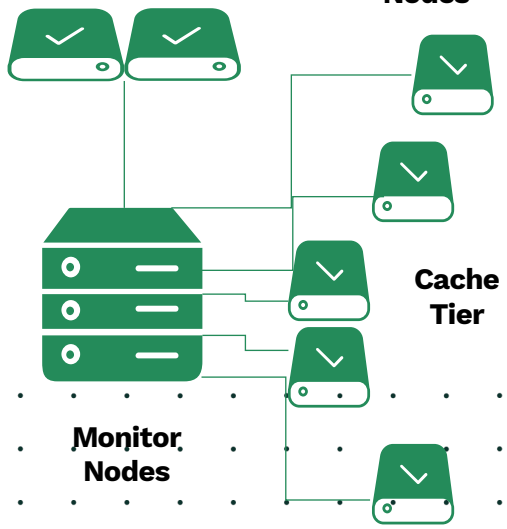
Primary HPC Storage

Ideal for small clusters (Ex: <= 250 nodes)



CephFS:

- Is a distributed file system with POSIX semantics
- Offers scale-out load-balanced active metadata servers and direct access to OSD nodes
- Cache Tier sized to working data set allows acceptable latency
- CephFS throughput scales with additional nodes



iRODS file virtualization layer:

- Reduced opex by automating workloads
- Eliminates storage silos
- Single pane for data management

iRODS and SUSE Enterprise Storage

iRODS
Data Grid

What is it?

- integrated Rules-Oriented Data System
- Highly flexible open source storage middleware
- Developed and supported by iRODS consortium (as opposed to a commercial entity)

How does it work?

- Placed between storage and applications
- Integrates multiple storage tiers

High level use cases

- Data management
- HPC storage

What does SUSE joining the consortium mean?

- SUSE Enterprise Storage tested in their lab; SUSE customers get support for iRODS connectivity
- SUSE will have a voice on iRODS features and development
- SUSE will be mentioned in iRODS materials

SUSE

Addressing the Challenges

Today's Complexity.

Meet Jin,
Data Scientist

Jin tests model
(local)

Jin writes model

Jin transforms model
(local) into a pipeline
compliant file

Jin uploads model to
AI platform

Meet Sasha,
AI/ML Ops

Pipeline execution

Run and Fix
pipeline until
success

Deployment to
Production

DevOps have no real control

Time T0 (Zero)

Time T1 (Week 2)

Time T2 (Week 4)

Time T3 (Unpredictable)

Time T4
(T3+1 day)

Challenges in getting an AI/ML project into production in a timely manner while meeting requirements

SUSE

Tomorrow's Automation.

Orchestration, collaboration and automation gets the AI/ML project into production in a timely manner while meeting requirements

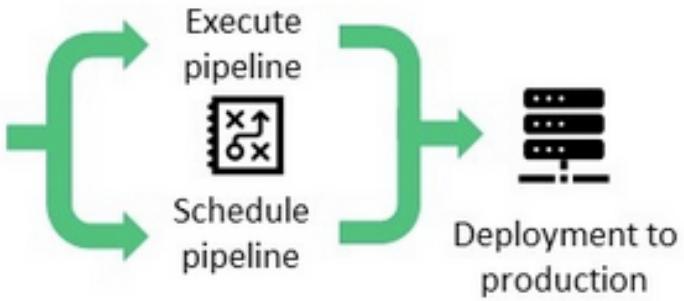
Jin tests model (local) and fixes errors

Jin writes model

Model is automatically uploaded to Git repository

Jin submits model to the **SUSE AI Orchestrator**

Sasha receives a notification and evaluates Jin's submission.



Time T0 (Zero)

Time T1 (Week 2)

Time T2 (T1 + 2/3 days)

Time T3
(T2 + pipeline execution time)

Time Reduction

SUSE

Tomorrow's Automation.

Jin tests model (local) and fixes errors

Jin writes model

Model is automatically uploaded to Git repository

Jin submits model to the **SUSE AI Orchestrator**

Dev

TensorFlow
beam
Kubeflow
argo
Apache Airflow
Google Cloud
Amazon SageMaker
Microsoft Azure

Test

Prod

Time T0 (Zero)

Time T1 (Week 2)

Time T2 (T1 + 2/3 days)

Time T3
(T2 + pipeline execution time)

Time Reduction

SUSE

Approach to AI/ML Orchestration

- **Quick start** with templates, optimized for specific environments (e.g., GPU accelerators) or customize to your needs
- **Configure multiple environments** for same pipelines – core to cloud – with one click
- **Pre-built automation workflow** removes complexity with freedom to change infrastructure later
- **Auto-pilot guidance** to drive the AI operator through entire deployment experience – deploy what you need
- **Reduced friction** for accelerated projects with a tested and maintained solution

SUSE AI Orchestrator

What is it?

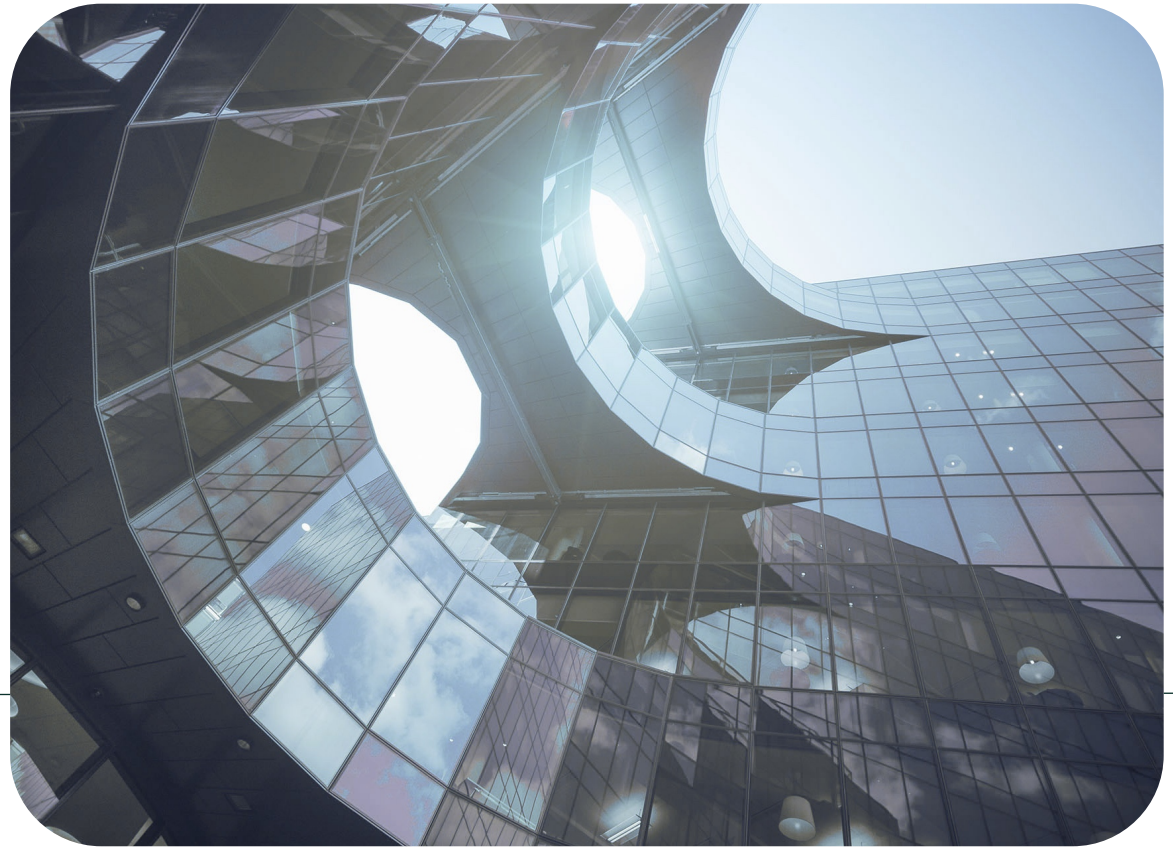
- A cloud-native tool that translates a data model into the execution steps of an AI platform pipeline or workflow in an automated way

How does it work?

- When the data scientist submits an execution request against the unmodified data model, the tool discovers and applies all the runtime options required for the chosen AI platform

Why use it?

- **Automates** pipeline or workflow across AI platforms
- Fosters **collaboration** between data scientists and AI operators
- **Monitor or deploy** an entire AI platform on-premise or in the cloud

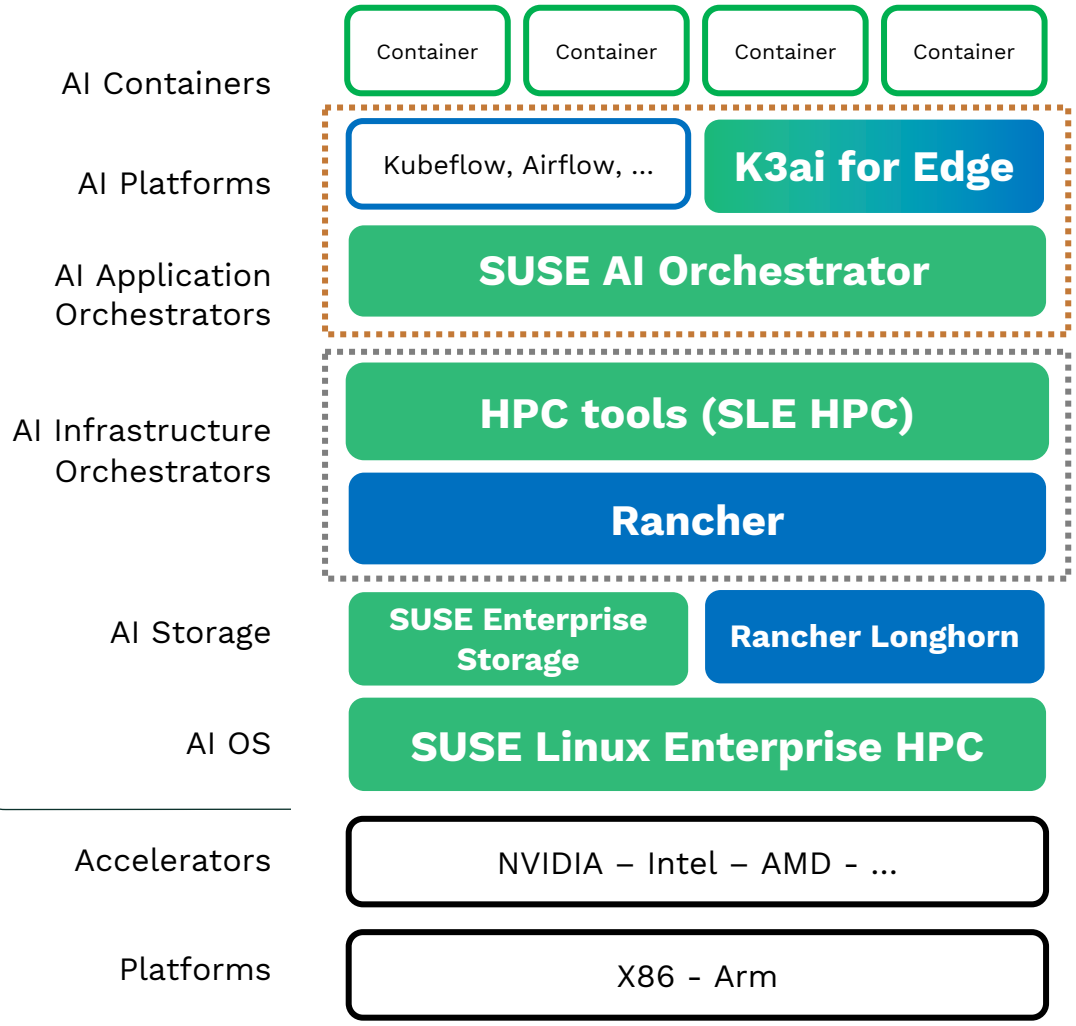


AI Infrastructure for Edge

- K3ai (https://docs.k3ai.in/) goal is to build an omni-comprehensive solution based on Rancher K3s and popular AI tools and platforms.
- Current version of k3ai supports:
  - NVIDIA GPU operator
  - Kubeflow pipelines
  - Kubeflow Full (WIP)
  - Argo Workflows (WIP)
  - Tensorflow serving
  - NVIDIA Triton Inference Server (WIP)
  - Seldon Inference Server (WIP)
- K3ai offers infrastructure for edge devices with full capability of a Kubernetes cluster
- Installed and redeployed with single command
- Supports HA and multi-node edge clusters
- Runs on Arm and x86

SUSE AI Stack

- Complete AI infrastructure stack that combines:
  - Server hardware
  - Hardware abstraction layers
  - Orchestration layers
  - AI development layers
  - Data science layers
- These layers seamlessly operate together



Learning More

Snapshot of AI Projects.

AI Orchestration

- Reduce AI/ML project time to production through automated workflows that can be changed as hardware changes

AI Infrastructure for Edge

- Ease deployment and management of AI/ML with software building blocks and guidance

High-Performance Computing Tools

- Provide a powerful platform and tools for data-intensive AI/ML workloads across core, cloud and edge

Discover More.

AI Orchestration

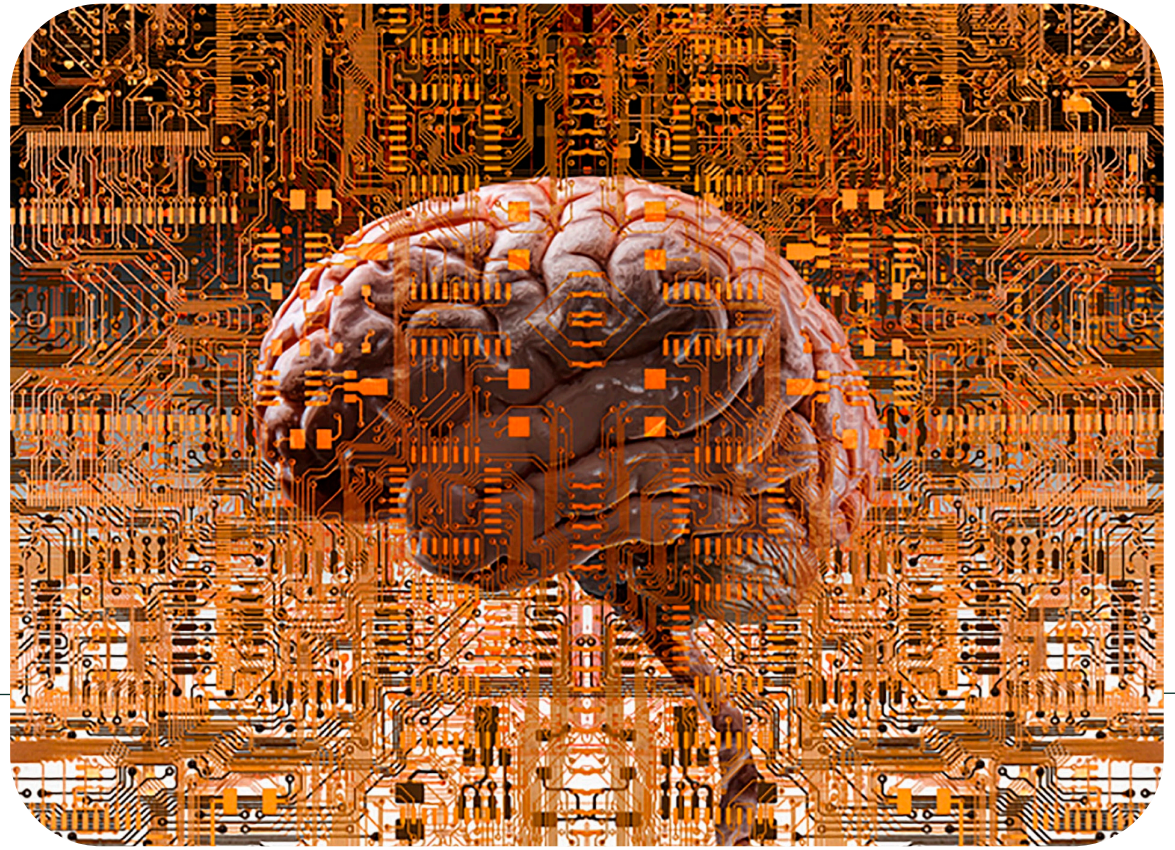
- Machine Learning pipelines for Kubeflow https://github.com/kubeflow/pipelines/pull/4278

AI Infrastructure for Edge

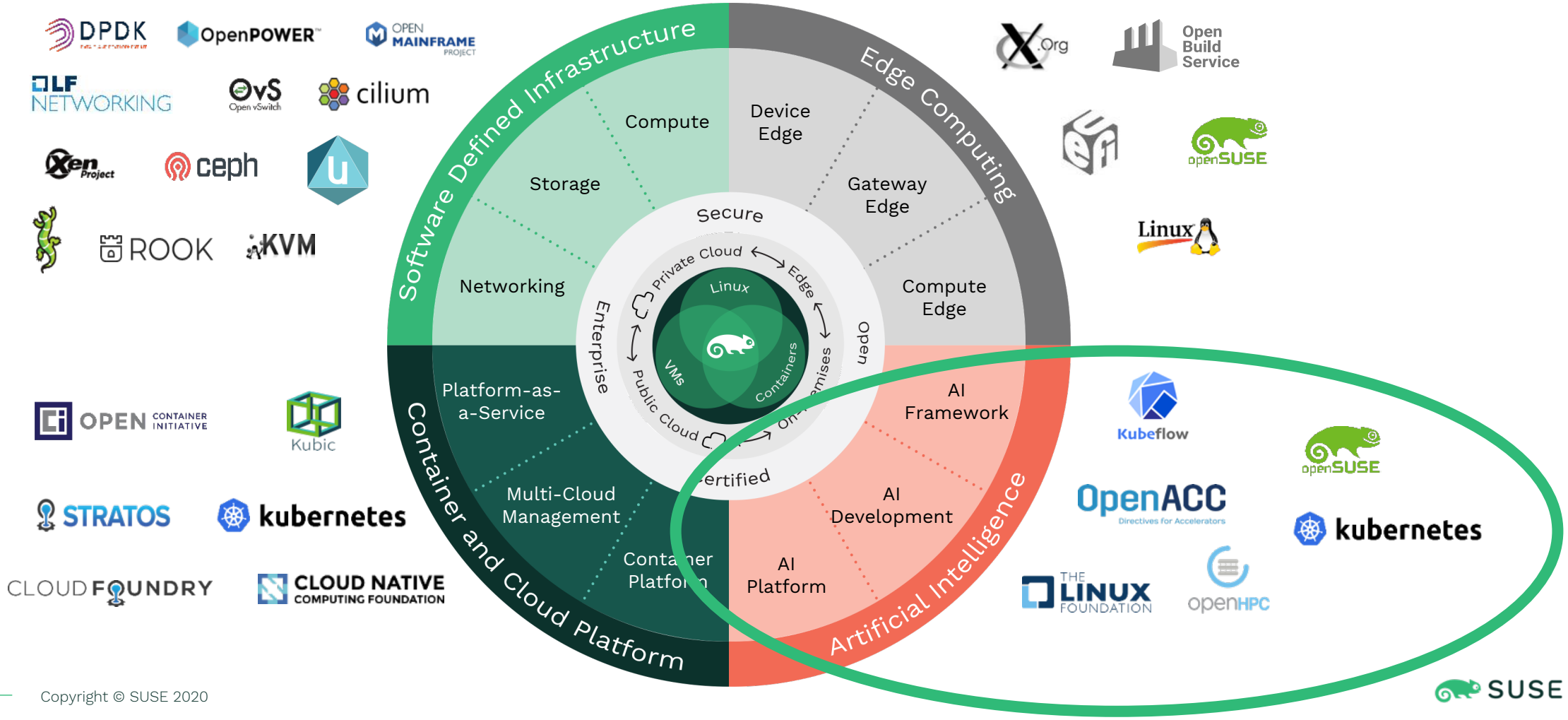
- K3ai is a guidepost for building and deploying optimized AI models https://docs.k3ai.in/

High-Performance Computing Tools

- SUSE Linux Enterprise HPC https://www.suse.com/products/server/hpc/



AI and Open Source Communities



SUSE

Thank you.

Jeff.Reser@suse.com

For more information, contact SUSE at:

+1 800 796 3700 (U.S./Canada)

+49 911 740 53-0 (Worldwide)

SUSE.com

SUSE