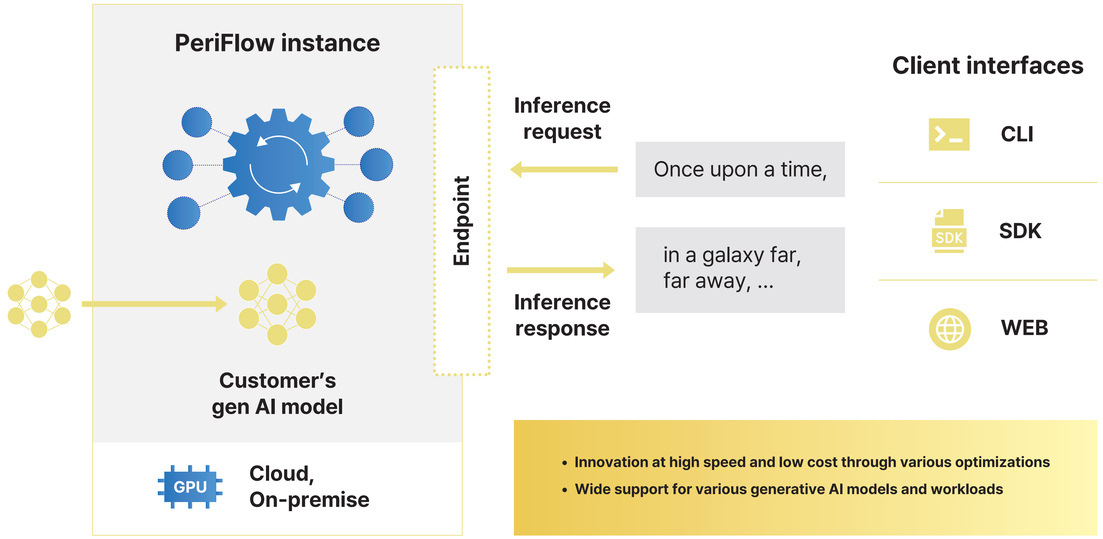
PeriFlow
Engine for serving generative AI models

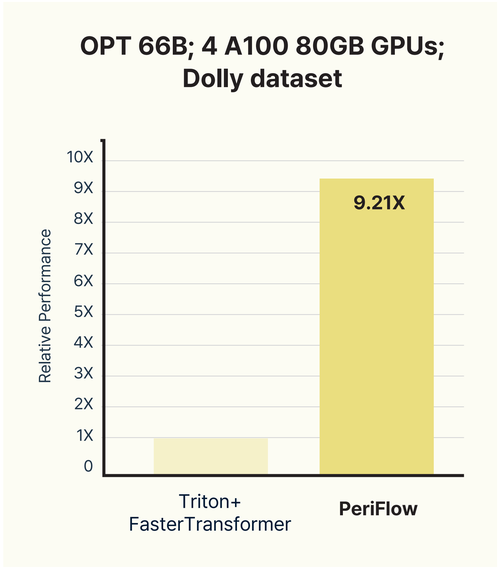


- Innovation at high speed and low cost through various optimizations
- Wide support for various generative AI models and workloads

PeriFlow
Engine for serving generative AI models

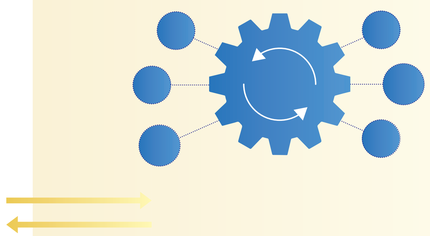
- 10x faster than Triton and FasterTransformer
- Reduce GPUs needed by 10x → reduce costs, improve reliability
- Wide support for various gen AI models and workloads
- Streamline gen AI inference optimization and deployment

(* actual gains depend on workloads)
(* patents)



Two ways to use PeriFlow :
PeriFlow Container and PeriFlow Cloud

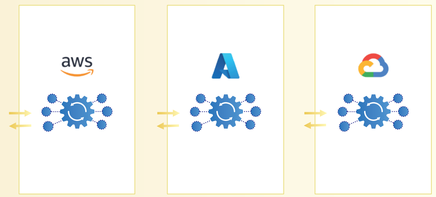
PeriFlow Container



Run PeriFlow in your own controlled environment

PeriFlow Container generally available

PeriFlow Cloud

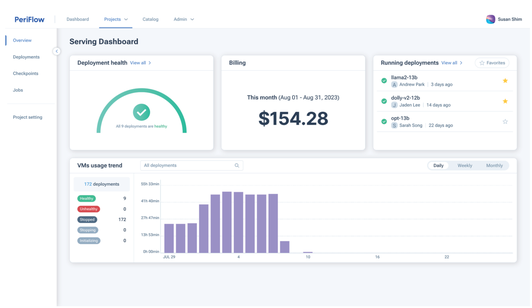


Run PeriFlow on autopilot

Public beta available

PeriFlow
Cloud interface

Effortless deployment



Automatic management