Contribution ID: **123** Type: **Oral Presentation**

Resource Balancing in AI/HPC Accelerated Solutions

Tuesday, 21 March 2023 09:50 (40 minutes)

Accelerating time to market, time to service and time to science through computing, the challenge we encountered today may not be in the speed of data processing. To streamline huge amount of data without blocking and queuing in an AI/HPC infrastructure design is crucial.

In this speech, a case study about resource balancing in an AI driven inspection and
decision systems from a world top notch smart manufacturer will be presented and will be showcased the design of the infrastructure with scale up/out upgrade path as the demands grow. The similar concept can also be employed in the problem solving for the other fields of AI/HPC applications like high-energy physics, nuclear and astrophysics.

To strengthen the weakest link of the chain in this particular case, the team examined and analyzed the problem from the manufacturing processes, the system with proper balancing infrastructure design between computing capacity and the performance of tier0 storage has been delivered with plug and play readiness in rack scale.

Last not least, a clear and present challenge and solutions in AI/HPC systems will also be quickly addressed at the end part of the session.

Presenter: Dr LYNN, Andrew (Super Micro)

Session Classification: Opening Ceremony & Keynote Speech I