Q3

The comparison between P100 and A100:

P100:

Extreme performance
Powering HPC, Deep Learning, and many more GPU Computing areas
NVLink™
NVIDIA's new high speed, high bandwidth interconnect for maximum application scalability
HBM2
Fast, high capacity, extremely efficient CoWoS (Chip-on-Wafer-on-Substrate) stacked memory architecture
Unified Memory, Compute Preemption, and New AI Algorithms
Significantly improved programming model and advanced AI software optimized for the Pascal architecture;
16nm FinFET
Enables more features, higher performance, and improved power efficiency

A100:

3rd generation Tensor Cor
New format TF32, 2.5x FP64 for HPC workloads, 20x INT8 for AI inference, and support for BF16 data format.
HBM2e GPU memory
Doubles memory capacity compared to the previous generation, with memory bandwidth of over 2TB per second.
MIG Technology
Each instance offers up to 7 isolated Multi Instance GPUs (MIG), each with 10 GB of RAM.
Special support for sparse models
For sparse matrix calculations (tensors with many zeros), provides a 2x compared to the previous generation.
3rd Generation NVLink and NVSwitch
upgraded network interconnect enabling GPU-to-GPU bandwidth of 600 GB/s.

CONFERENCE:

1. https://www.run.ai/guides/nvidia-a100
2. https://www.nvidia.com/en-us/data-center/a100/
3. https://images.nvidia.com/content/pdf/tesla/whitepaper/pascal-architecture-whitepaper.pdf