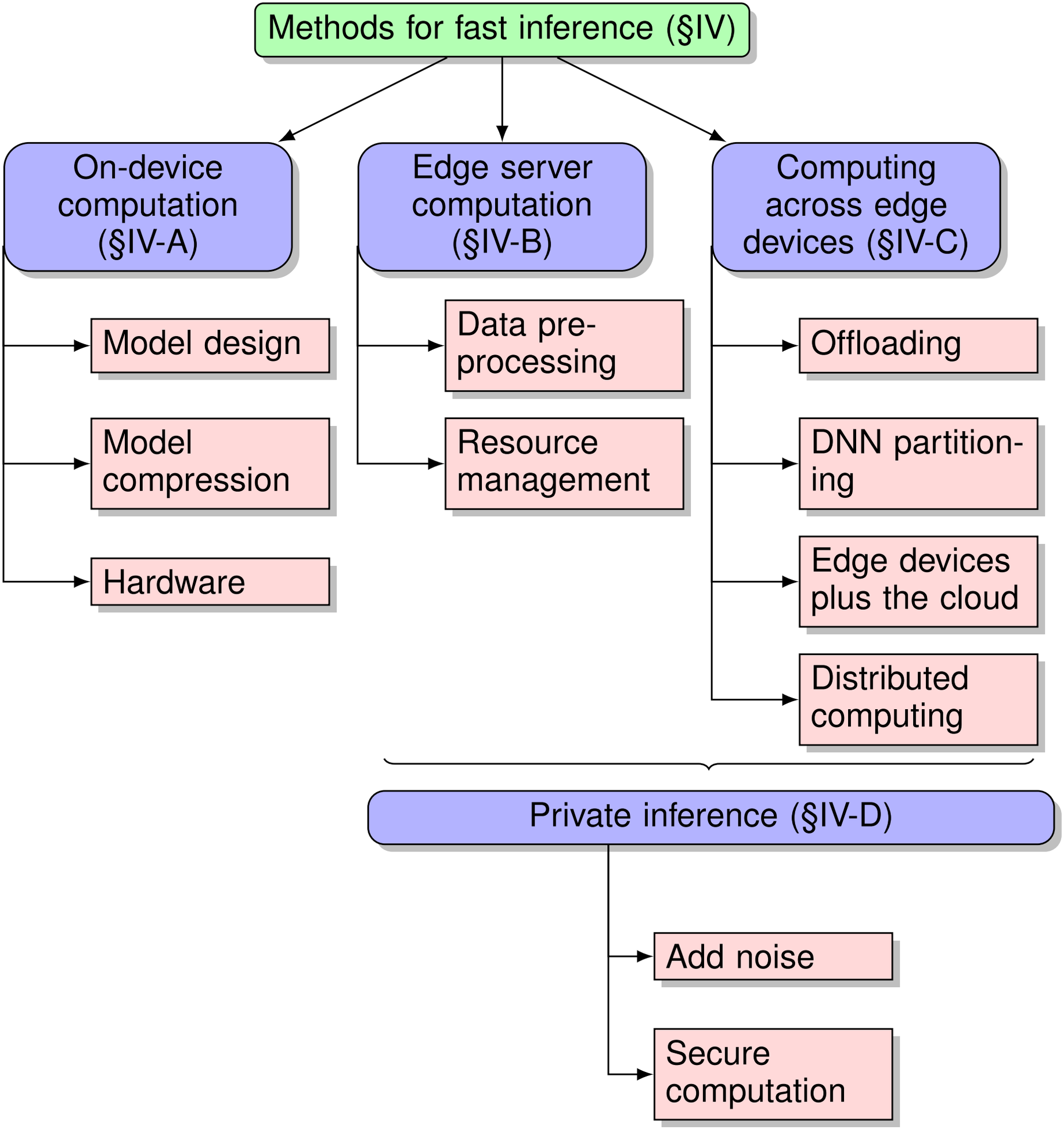
Methods for fast inference (§IV)
On-device computation (§IV-A)
Edge server computation (§IV-B)
Computing across edge devices (§IV-C)
Model design
Model compression
Hardware
Data pre-processing
Resource management
Offloading
DNN partition-ing
Edge devices plus the cloud
Distributed computing
Private inference (§IV-D)
Add noise
Secure computation