Neural Network Analysis for Next-Generation Video Compression Standard

Ted Hsieh, Wei-Jung Chien, Vadim Seregin, Marta Karczewicz

Qualcomm

Agenda

- Possible locations of neural network(NN) processor for video coding processing
- Complexity Analysis cases
  - IntraNN
  - ALF-NN
- Suggestions

Possible locations of NN processor for video coding processing

- Video coding processing unit and NN processor are on two separate SoC
  - Video codec and NN processor are on two different dies
    - i.e. NN processor is outside of traditional video codec core
    - NN processor could be utilized to execute general AI functions
      - Commercial GPU/NPU
    - Data exchange between video codec and NN processor can be through
      - PCIe / DDR or HBM

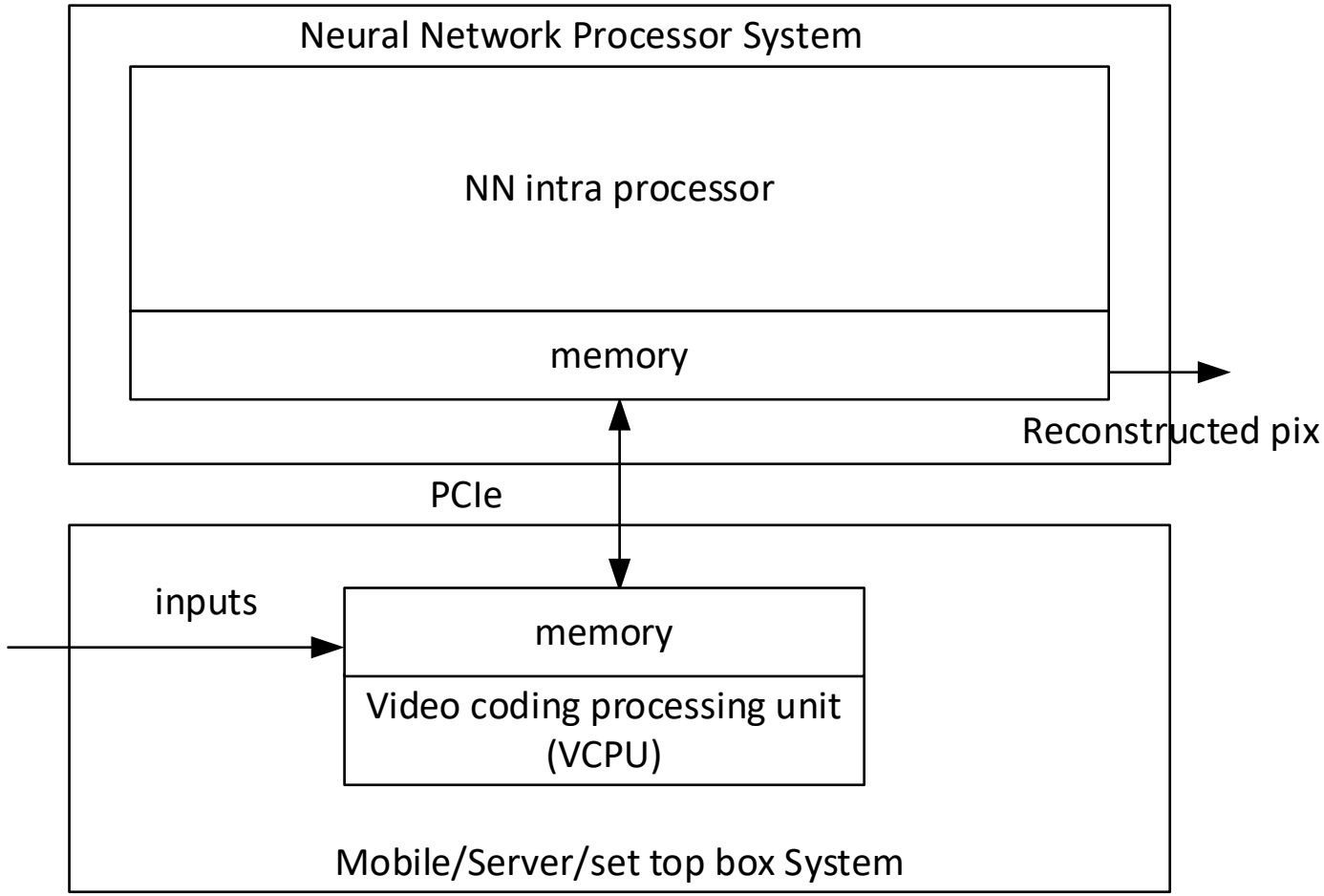
- Video coding processing unit and NN processor are in the same SoC
  - NN processor will reside in the same die of video codec
    - i.e. NN processor is part of video codec core
    - NN processor is dedicated to execute video's particular tools/functions
      - In-house design
    - Data exchange between video codec and NN processor could be through
      - Direct Bus / DDR / Local Mem

Complexity Analysis

- Two tools have been investigated
  - Intra NN
    - Separate SoC for IntraNN and video codec processing unit(VCPU)
    - Same SoC for IntraNN and VCPU
  - ALF-NN
    - Separate SoC for ALF-NN and VCPU
    - Same SoC for ALF-NN and VCPU

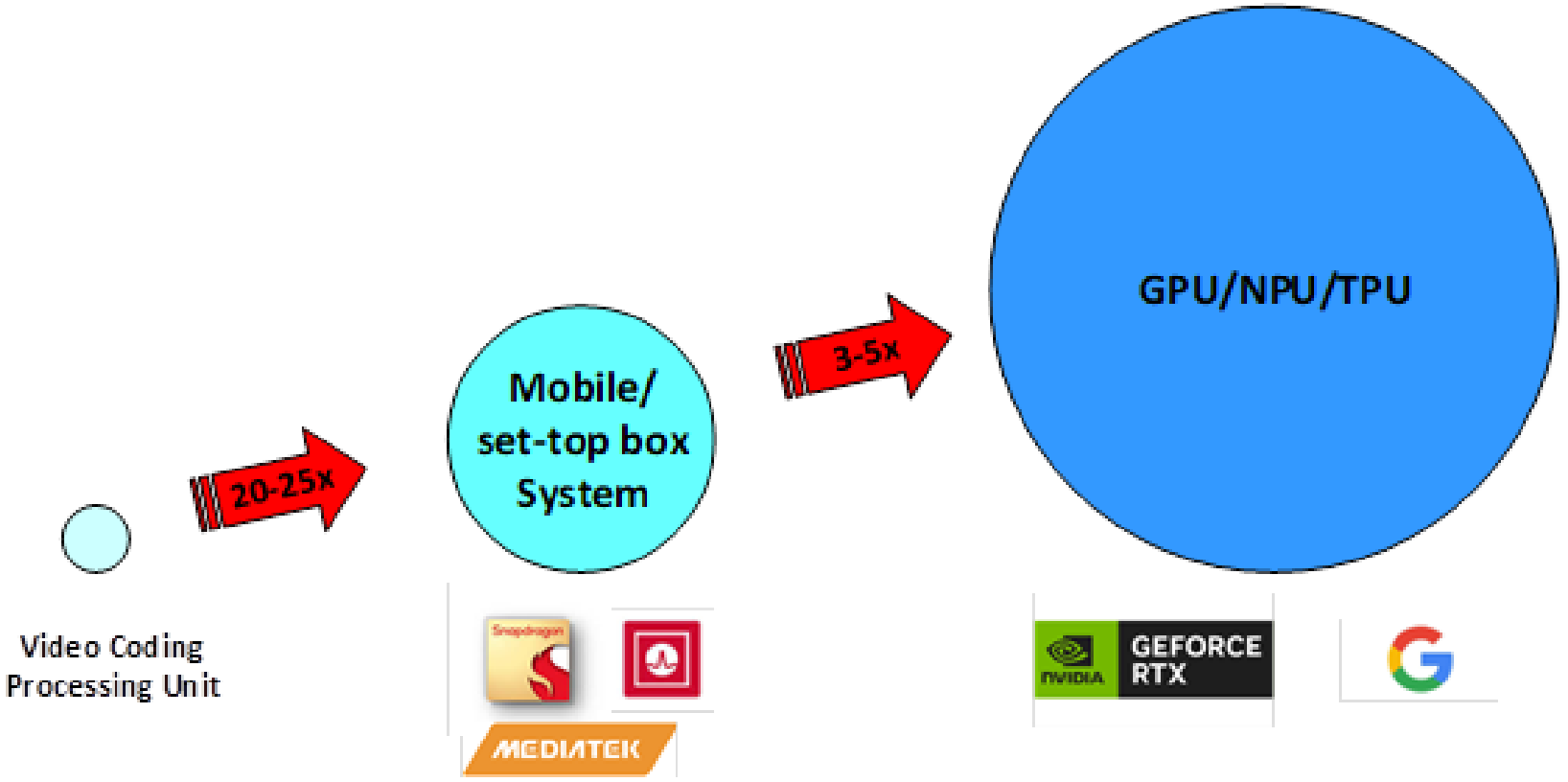
IntraNN

- VCPU and NN processor are on two separate SoCs
  - Simplified system block diagram



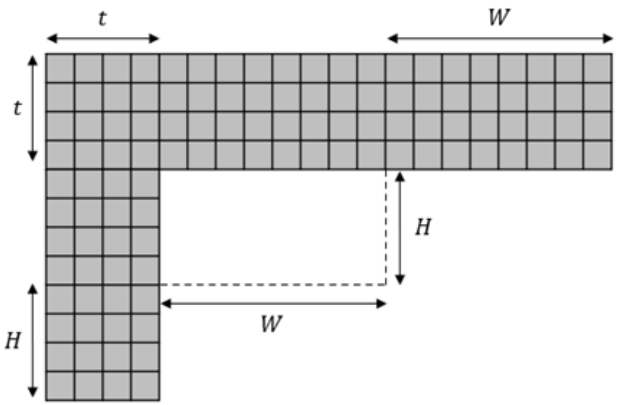
IntraNN

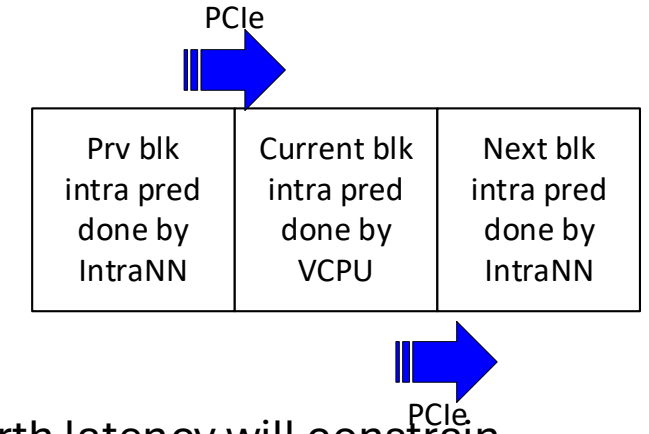
- VCPU and NN processor are on <u>two separate SoCs</u>
  - Complexity/area:
    - The estimated areas ratio of separate GPU/NPU SoC roughly can be from 60x to 125x of area of traditional video codec ASIC



IntraNN

- VCPU and NN processor are on <u>two separate SoCs</u>
  - Communication/ Data transfer
    - PCIe is used as communication channel for commands, signals, data, et
      - High data transmit rate: ~8GB/s
      - Latency
  - The worst latency of IntraNN prediction example
    - Dependency paths
      - Current block's intra prediction will be done by traditional VCPU intra prediction, but need to wait for previous block's IntraNN's rec sample through PCIe
      - Next block's intra prediction will be done by IntraNN, which also need to wait for current block's VCPU's rec sample s through PCIe

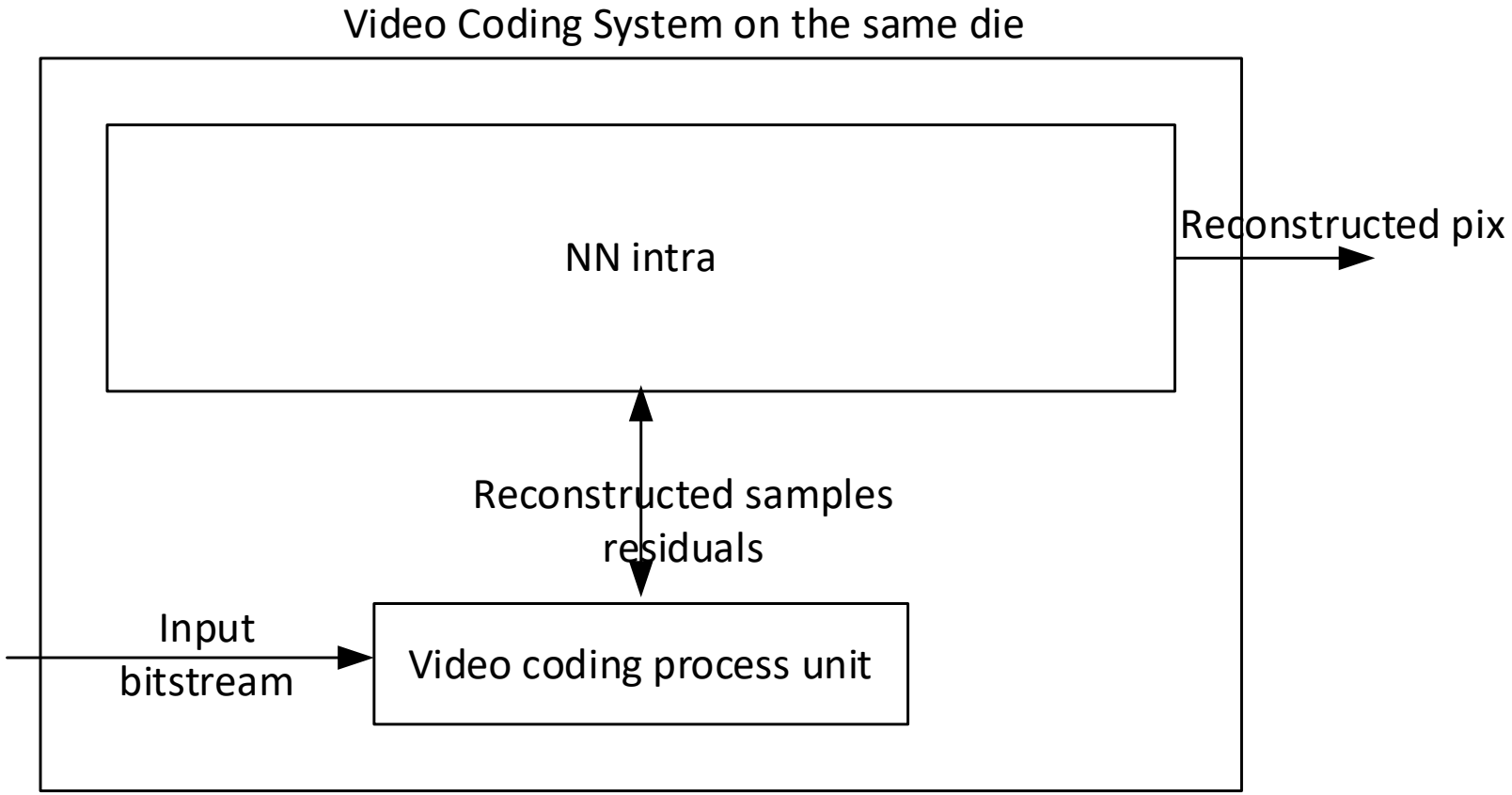




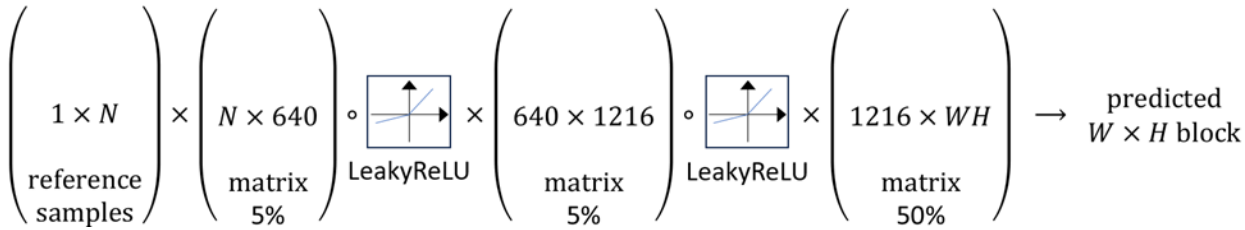
- Summary
  - The factors lead to an impractical system of the separate IntraNN SoC
    - High cost/complexity: external GPU/NPU's large area and $
    - Latency: even PCIe can provide high data transmit rate, but the back and forth latency will constrain external IntraNN processor for the real time applications
    - Sequential process for block by block, less efficient on GPU utilization and data transfer

IntraNN

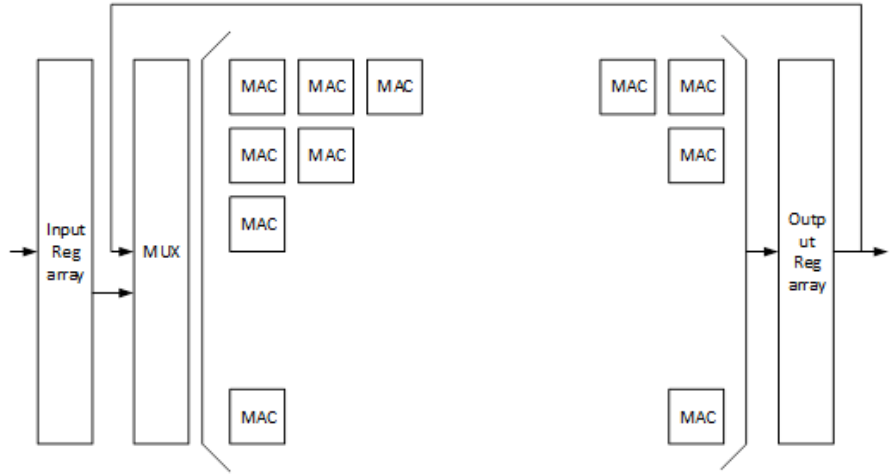
- VCPU and NN processor are on same SoC
  - Simplified system block diagram



IntraNN



- VCPU and NN processor are on same SoC
  - Complexity/area:
    - Architecture design:
      - Based on shared-MAC arch we choose the supper set of MAC
        - can handle worst case: intra4x4's intra prediction most computation intense layer
      - Hypothesis is that model may reduce the number of MAC needed if sparsity of IntraNN is utilized
        - Supper set layer of MAC: 640x1216x5%= 38912 MACs, w/ 5% Sparsity /weights are non-zero
    - Area comparison
      - Sparsity
        - ~2X of VCPU (MAC only, no memory, no control)
        - if add in MUX for selection could become 1000X
        - Need to consider 6 shapes x 3 layers
      - Normal(non-sparsity)
        - 640x1216 = 778,240 MACs
        - ~40X of VCPU
    - Timing: for intra4x4 worst case
      - Total 5 layer = 3 Matrix layer + 1 input layer + 1 layer
      - Assume 1 cycle finish 1 layer computation, 5 cycle needed per 4x4



  - Summary
    - Major factors to make IntraNN less attractive for in-house design on same SoC:
      - Area cost is too high
      - Due to data dependency, system running frequency need to go higher if larger resolution, or higher frame rate
      - Power consumption(proportional to area and running freq)

ALF-NN

- VCPU and NN processor are on two separate SoCs
  - Complexity/area: similar analysis as IntraNN
  - Data transfer
    - Bandwidth: typical external GPU/NPU date transfer rate is high, so it's enough to handle parameters, input neighboring samples and output filtered samples
    - Latency path:
      - since commercial GPU/NPU data rate is high, the filtered recon samples write out should be fast enough
      - Maybe bottleneck will be the user case of super high frame rate
      - The last couple blocks of previous frame just finished the ALF-NN filtering and waiting to transmit to memory, but the current frame's first inter CU's MV has already started decoding and its MVs refer to the previous last couple blocks for inter prediction.

- Summary
  - The factors could impact the system of separate ALF-NN SoC
    - High cost and power of external GPU/NPU
    - Task priority for external GPU/NPU utilized by video applications

ALF-NN

- VCPU and NN processor are on same SoC
  - Complexity/area:
    - the area of ALF-NN is close to video codec's area
  - Timing:
    - Latency path:
      - 46 layers delay + memory write(ALF-NN -> mem) latency + memory read(mem -> video codec) latency

      => latency varies (depend on system), so it should be resolved by higher running frequency of ALF-NN; or

      => dedicate frame buffer between ALF-NN and video codec core => higher cost
  - Summary:
    - for above solutions, either running ALF-NN at higher frequency; or dedicated frame buffer
      - Both will result in more power consumption or higher chip area

Conclusion

- IntraNN
  - On separate SoC
    - Except high cost of NN, the sequential communication latency between VCPU SoC and IntraNN SoC, will make it impossible for the real-time video decoding
  - On same SoC(same chip)
    - To meet the requirement of worst case for intra4x4 decoding, cycle constraints of predictors generation result in high area cost, smaller but hardware need to support to finish 1 layer per 1 cycle
      - Sparsity of parameters does not help since more control/mux are needed for a MAC-shared architecture
    - For decoding higher resolution or higher frame rate, higher running frequency is needed, then power consumption also increases proportionally. It is not recommended to utilize IntraNN on the same chip.
- ALF-NN
  - On separate SoC
    - The higher cost and power for ALF-NN need to be considered for a separate SoC
  - On same SoC(same chip)
    - Also higher cost and power for its solutions(higher running frequency, dedicate shared memory), for resolving the latency issue.
- Recommendation:
  - Lots of imbalance requirements and different characteristics exist between VCPU and NN processor, like running frequency, data bandwidth, latency and math capability, so, instead of bridging the hybrid systems, maybe invest more efforts for an end-to-end NN video coding system developments