Update on Ara

08/07/2022

Matteo Perotti
Matheus Cavalcante
Nils Wistoff
Gianmarco Ottavi

Professor Luca Benini
Integrated Systems Laboratory
ETH Zürich

Summary

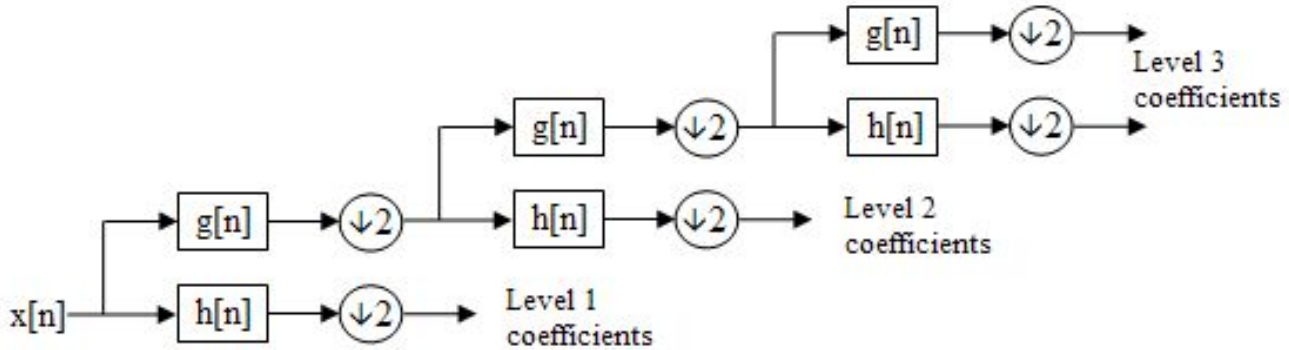
- **ASAP conference - Presentation**
  - Prepare slides
  - Record presentation
- **New benchmark**
  - DWT
- **Performance analysis**
- **HW**
  - FP reductions integration

Benchmarks - Analysis ongoing

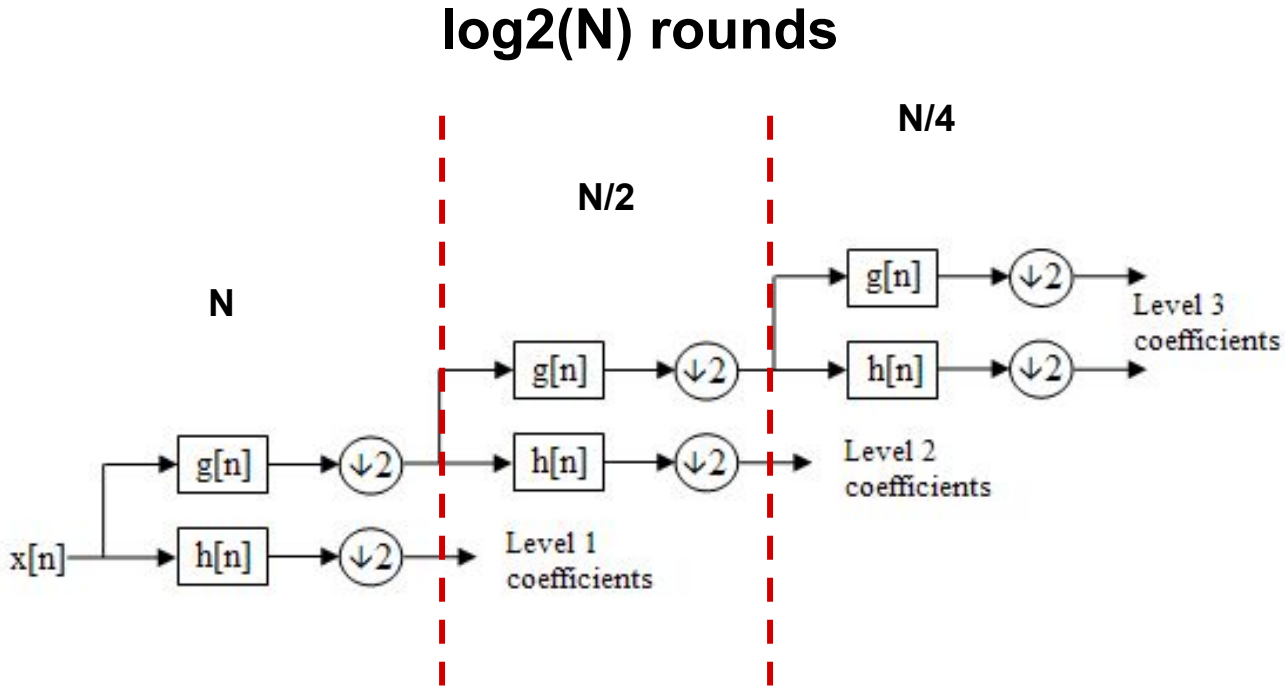
- `[i,f]matmul` - crucial kernel
- `[i,f]conv2d` - vslides
- `roi_align`
- `jacobi2d` - stencil, misaligned accesses
- `axpy` - mem bound
- `spmv` - indexed mem ops
- `dropout` - memory bound
- `dwt` - segmented or strided
- `fft` - segmented, masked permutations
- `softmax` - fpred, fpdivisions
- `float cos, log, exp`
- `memcpy, strncmp, strncpy`
- `[i,f]dotp` - reductions
- `pathfinder`

Benchmarks - DWT

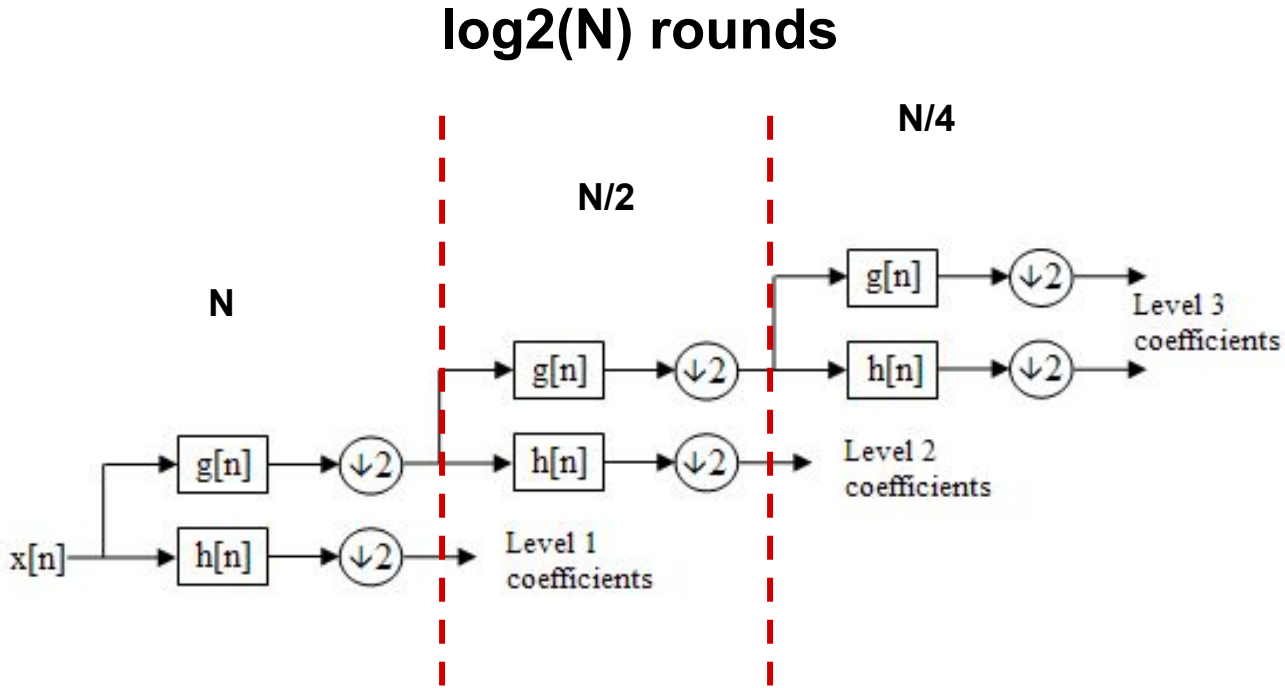
log2(N) rounds



Benchmarks - DWT



Benchmarks - DWT



First implementation: 2 coefficients/filter!

Benchmarks - DWT

- **DWT** (float32) - Memory bound
- **Downsampling**
  - ✖ Segmented mem ops
  - ✓ Strided mem ops
    - Intrinsic BW limitation
- DWT 512 samples
  - `Performance on max: 41% (48% only first iteration)`
- Improvement over scalar
  - 8x

Further

- Complete **performance analysis** for our benchmark pool
- Add **softmax (+exp, log, cos)**
- Add **FP reductions**
  - Solve timing criticalities
- For each benchmark, **report analysis**