MAC Design: Standard Floating-Point vs. Hybrid Logarithmic 4-bit
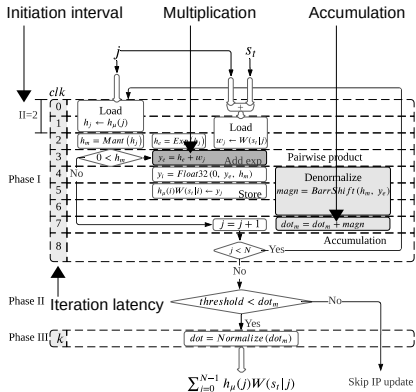

Standard floating-point

Initiation interval — **Multiplication** — **Accumulation**

j s_t

clk
0–9

Load
$h_j \leftarrow h_\mu(j)$

Load
$w_j \leftarrow W(s_t|j)$
$w_j = Float32(w_j)$

$0 < h_j$

Float multiply
$y_j = h_j w_j$

$h_\mu(j)W(s_t|j) \leftarrow y_j$ Store

Float adder
$dot = dot + y_j$

II=10

No

$j = j + 1$

$j < N$ Yes

No

$\sum_{j=0}^{N-1} h_\mu(j)W(s_t|j)$
Done

Iteration latency
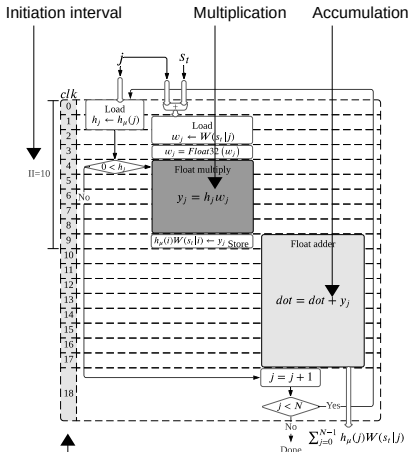
$L = (N-1)\ II + IL$

$L_{f32} = 10N + 9$

Hybrid logarithmic 4-bit

Initiation interval — **Multiplication** — **Accumulation**

j s_t

clk

II=2

Load
$h_j \leftarrow h_\mu(j)$

Load
$w_j \leftarrow W(s_t|j)$

$h_m = Mant(h_j)$ $h_e = Exp(h_j)$

$0 < h_m$ $y_e = h_e + w_j$ Add exp

Pairwise product

$y_j = Float32(0,\ y_e,\ h_m)$

Denormalize
$magn = BarrShift(h_m, y_e)$

$h_\mu(j)W(s_t|j) \leftarrow y_j$ Store

$dot_m = dot_m + magn$

Accumulation

Phase I

No

$j = j + 1$

$j < N$ Yes

No

Phase II   Iteration latency

$threshold < dot_m$ — No

Yes

Phase III k

$dot = Normalize(dot_m)$

Skip IP update

$\sum_{j=0}^{N-1} h_\mu(j)W(s_t|j)$

Iteration latency

$L_{log} = 2N+7$