Imperial College London

Low-latency option pricing using systolic binomial trees

Aryan Tavakkoli and David Thomas

Contributions

In pricing American options using the binomial-tree model:

- Proposed a spatial parallel hardware architecture
  - to provide low latency solutions
  - by exploiting modern large FPGAs
- Exploited the systolic nature of data-flow in a parallel binomial-tree model to
  - decrease latency and
  - maximize throughput
- Implemented a fixed-point accelerator
  - showing it provides sufficient accuracy for practical use
  - with speedups over previous approaches:
    - 65X in latency, 9X throughput

Scope

- Binomial-tree option pricing model
- Hardware architecture
- Implementation
- Results

Financial Options

- A financial option
  - A derivative financial instrument
  - that gives the investor the right
    - and not the obligation
  - to buy (*call* options) or sell (*put* options) an underlying asset
  - at an agreed-upon price (*strike price*)
  - by a certain date or at a specific date (*expiry date*)
- Example
  - A call option on Google's stock:
  - stock currently worth of $100
  - 1-year contract
  - strike price: $120
  - Next year, if Google's stock is:
    - $125 => exercised: pay-off: $5
    - $100 => not exercised: pay-off: $0

Options

- Option price
  - Pay-off is always non-negative
  - Options have values
- Option styles
  - Option types fall into different styles (families)
  - Examples:
    - American: can be exercised at any time until expiry date
    - European: can only be exercised at expiry date

Option Pricing: Parameters

- Main option parameters:
  - Spot price: S_0
  - Strike price: k
  - Time period: T
  - Volatility: σ
  - Interest rate: r

Option Pricing: Speed

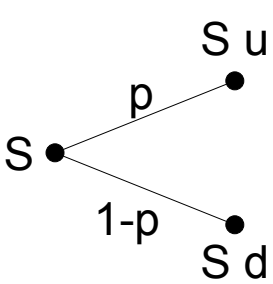
- Binomial-tree pricing model for American options
  - Pricing a single option using the binomial-tree model, on previous works:
    - CPUs: 1 - 100 *msecs*
    - GPUs: 0.1 - 1 *msecs*
    - FPGAs: 0.1 - 10 *msecs*
  - Practical use: up to 100s-1000s concurrent pricing
    - In large pricing baskets, or embedded Monte Carlo simulations
  - 1 second is significant for option evaluation competitors

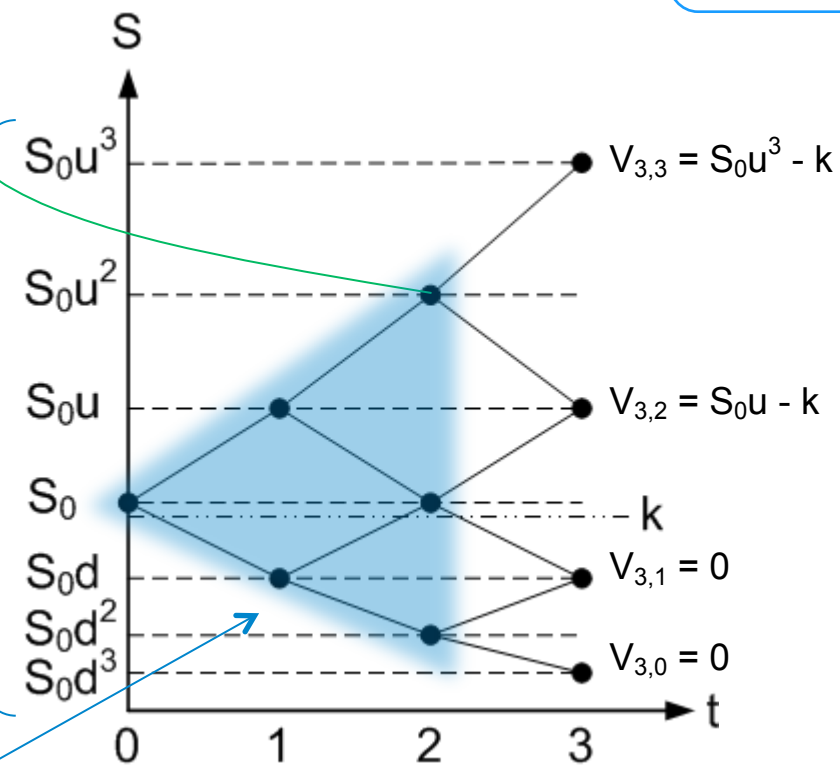
Scope

- Binomial-tree option pricing model
- Hardware architecture
- Implementation
- Results

Binomial-tree option pricing model

$$V_{2,2} = \max\{W_u V_{3,3} + W_d V_{3,2}, \max\{0, S_0 u^2 - k\}\}$$





asset price

$2n+1$ price levels

$2n+1$ exercise constants (B values)

call option, $n=3$ time steps

$n(n+1)/2$ node evaluations

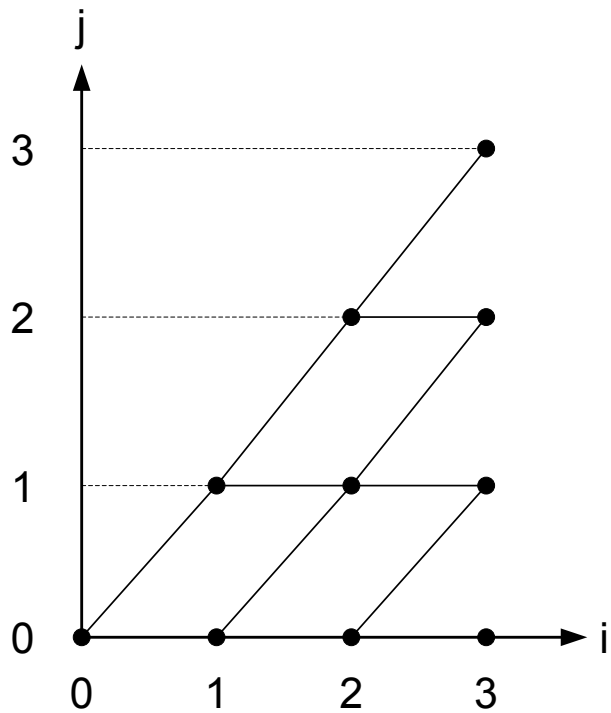
time

Scope

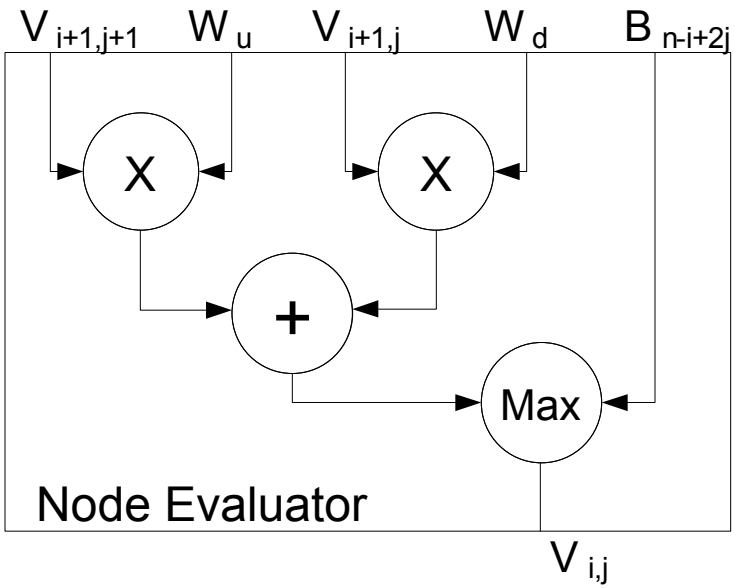
- Binomial-tree option pricing model
- Hardware architecture
  - Node Evaluators
  - Constant propagation mechanism
  - Systolic Array of Cells
  - Pipelined pricing concurrency
  - Performance model
- Implementation
- Results

Node Evaluator

$V_{i,j} = \begin{cases} B_{2j}, & \text{when } i=n \\ \max\{W_u V_{i+1,j+1} + W_d V_{i+1,j},\ B_{n-i+2j}\} & \text{otherwise} \end{cases}$

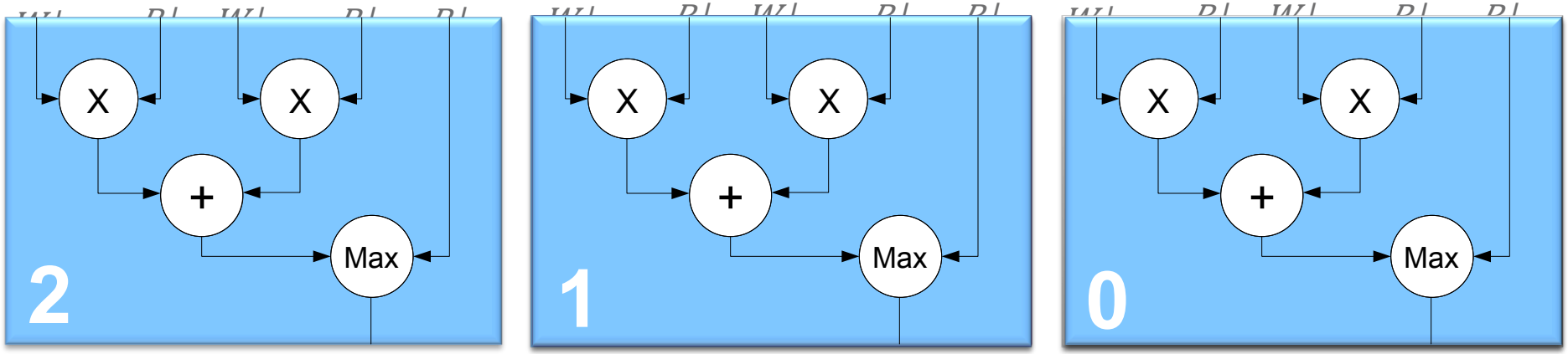


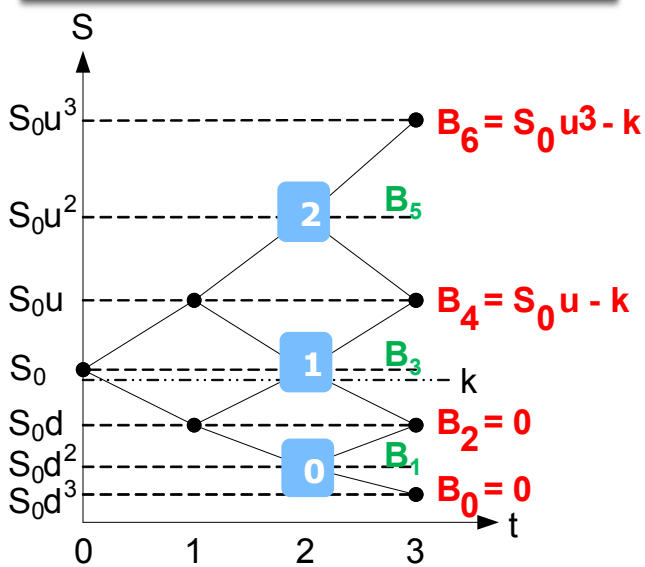
$n=3$



Data Flow

Pricing step #1

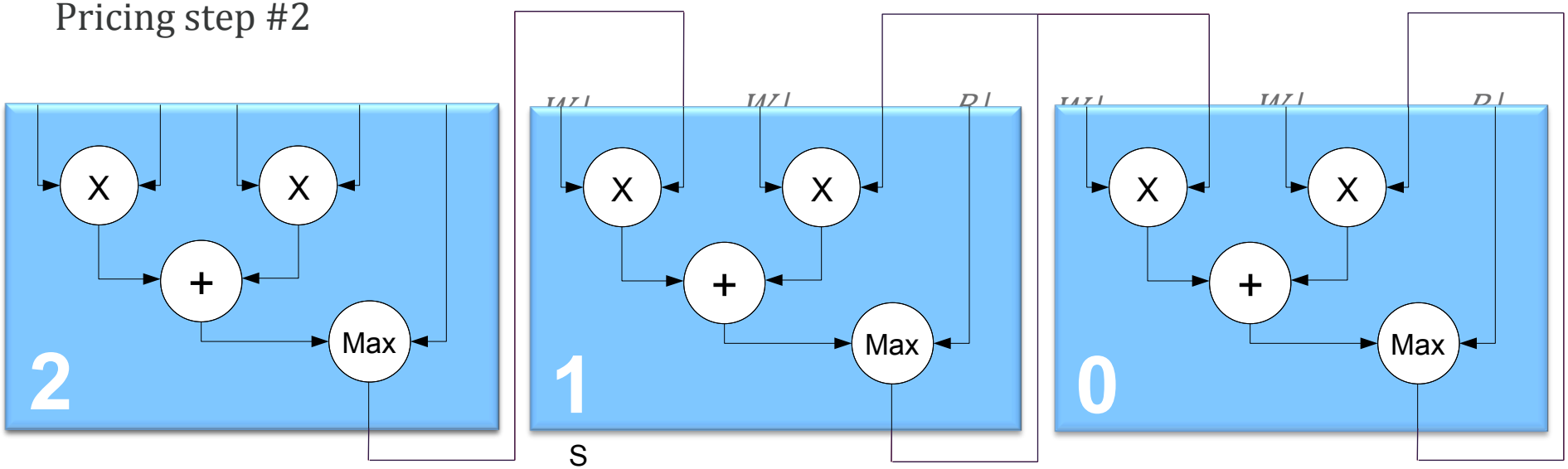


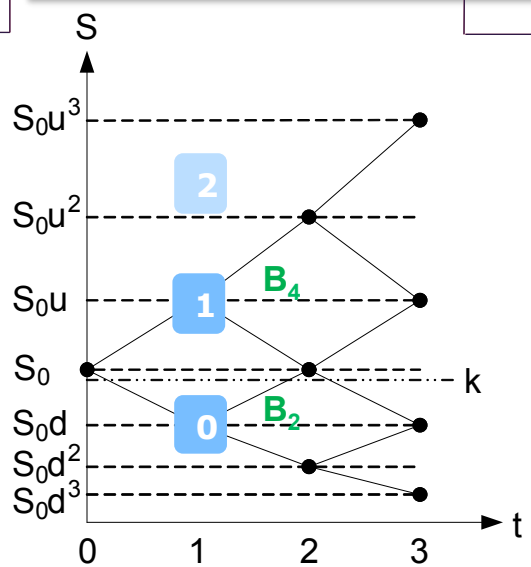


$n=3$

Data Flow

Pricing step #2

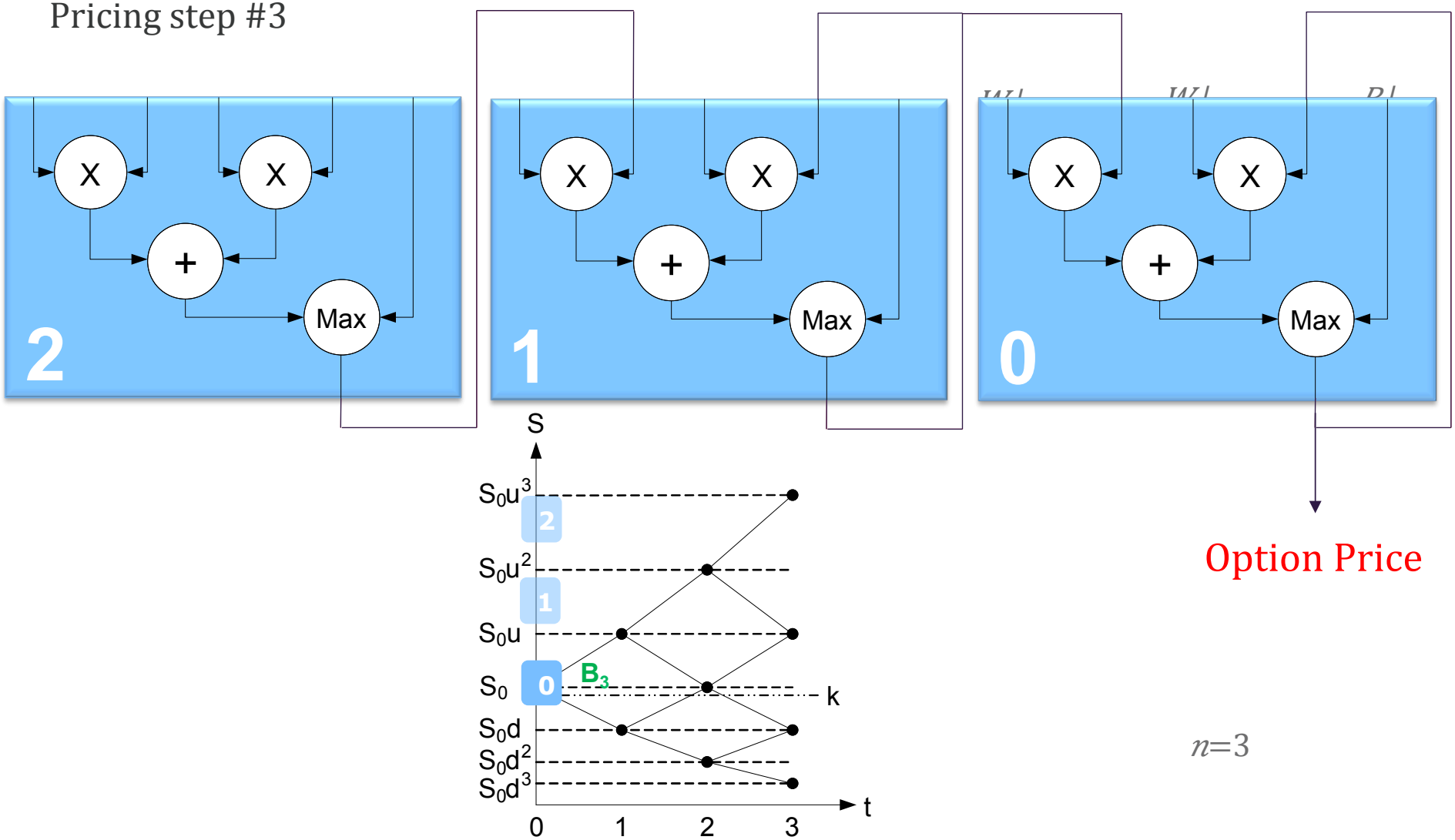




$n=3$

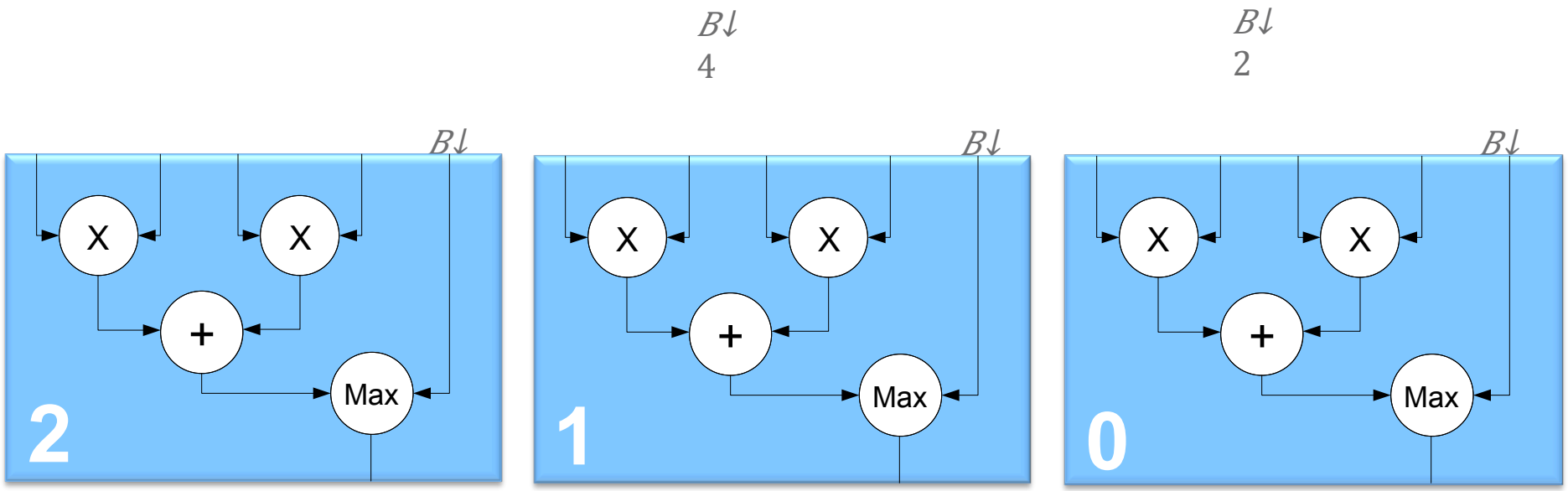
Data Flow

Pricing step #3



Constant Propagation

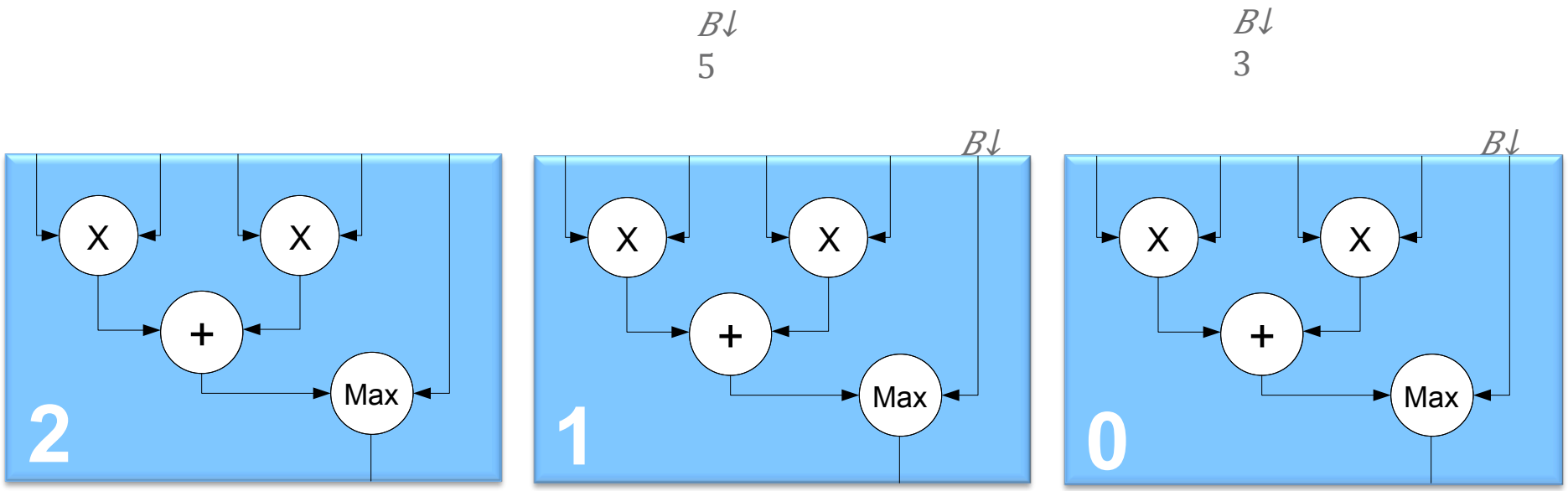
Pricing step #1



$n=3$

Constant Propagation

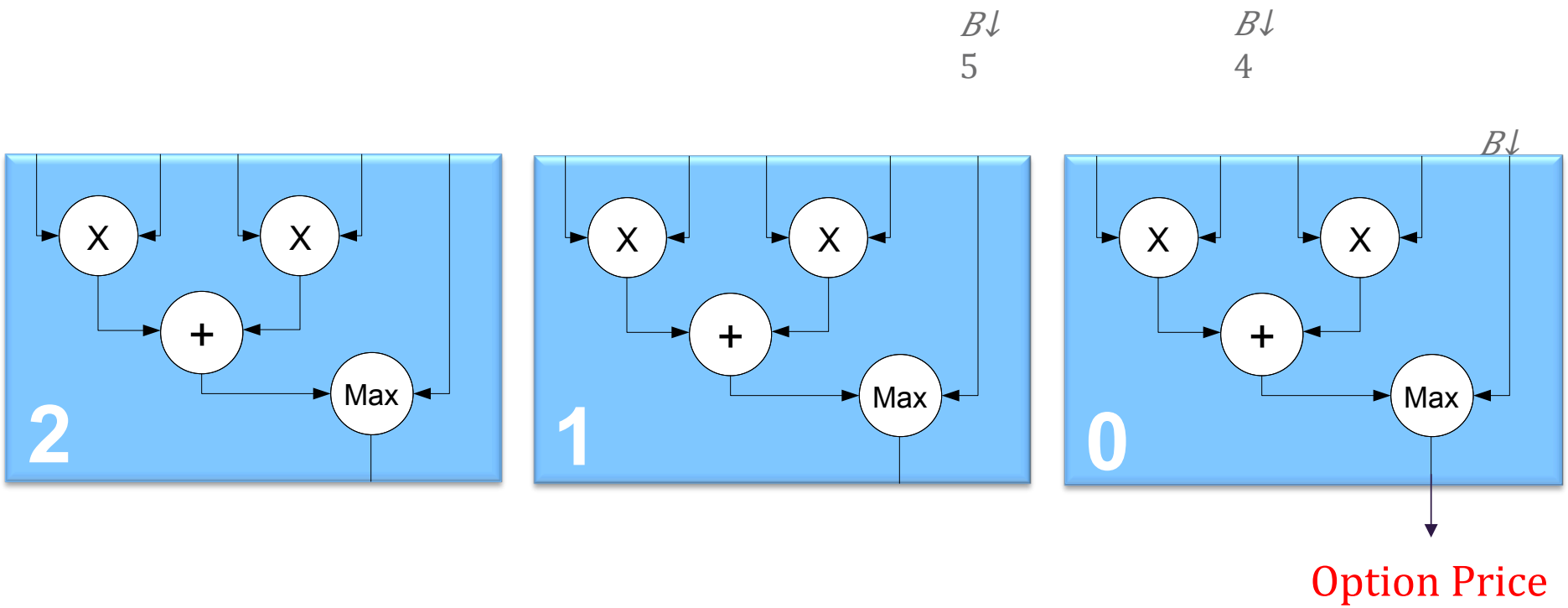
Pricing step #2



$n=3$

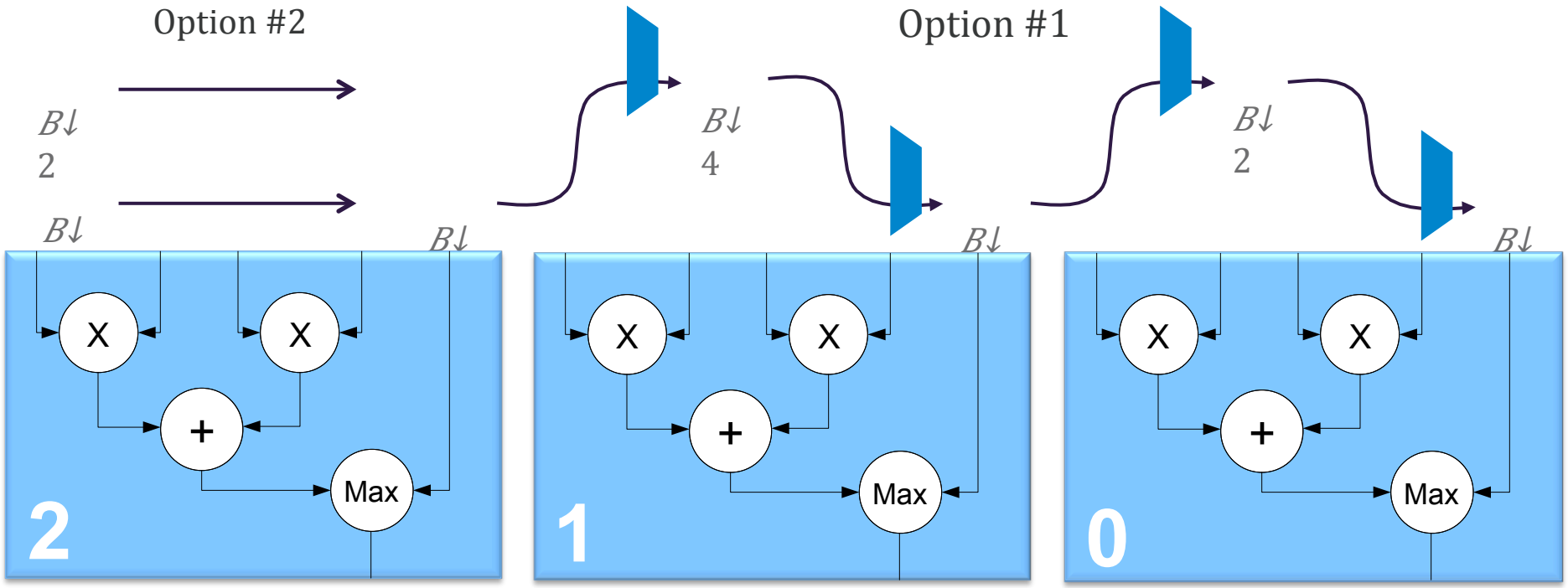
Constant Propagation

Pricing step #3



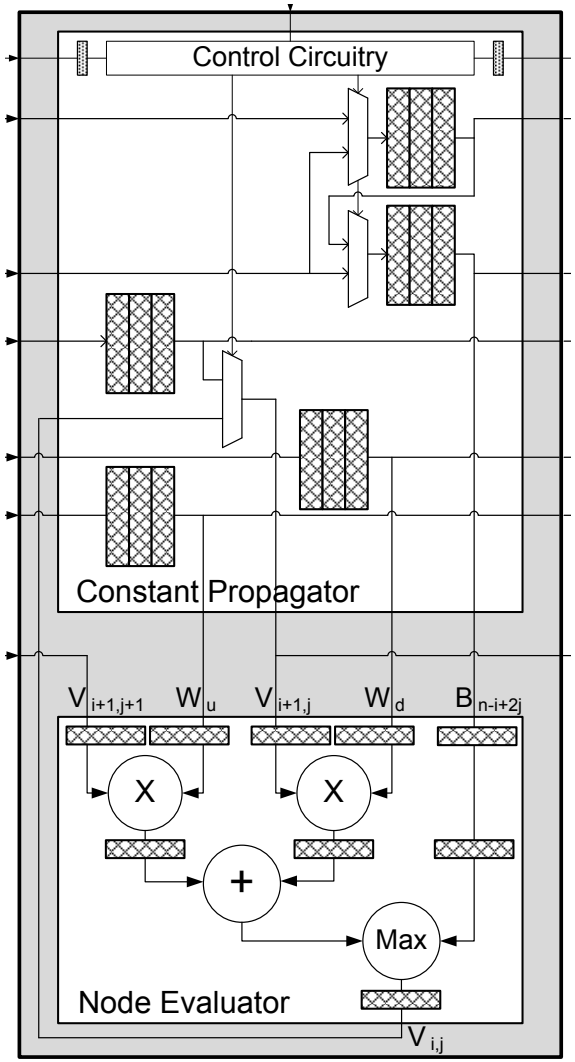
n=3

Constant Propagation



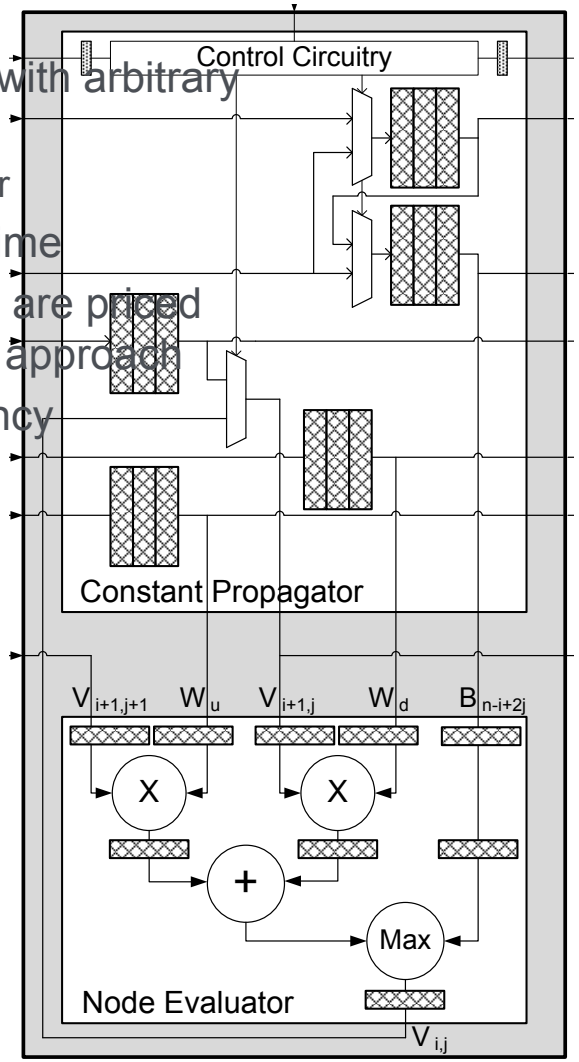
$n=3$

Systolic Cell



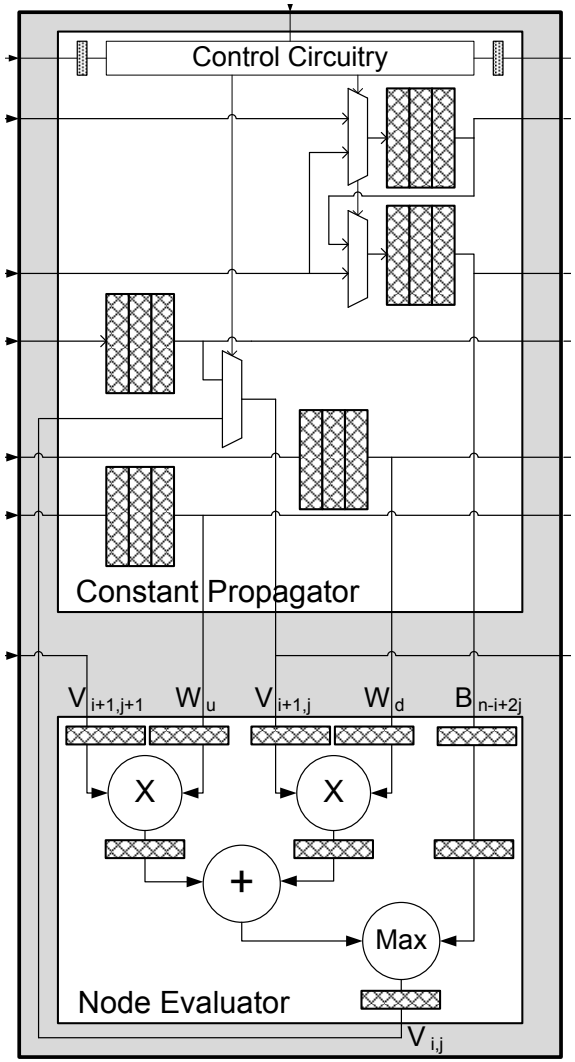
Concurrent Option Pricing

- Systolic Cell is pipelined with arbitrary number of stages
  - *m*: concurrency parameter
- Customised as compile-time
- An *option set* (*m* options) are priced concurrently in a C-slow approach
- *m*=1 gives minimum latency
- In the figure: *m*=3



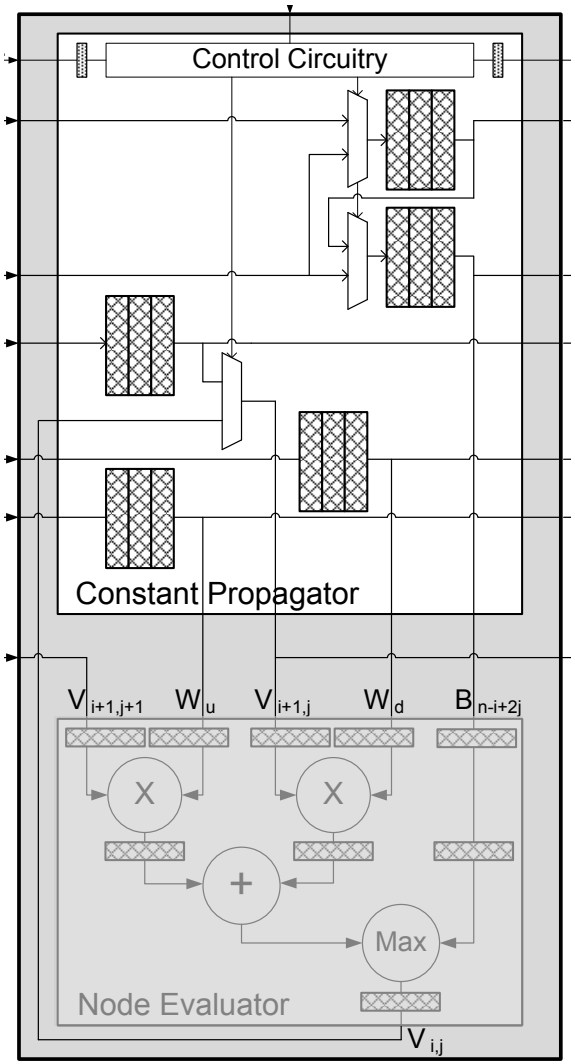
Systolic Cell: Modes of Operation

- **Evaluation mode**
  - Node Evaluator:
    - Active
  - Constant Propagator:
    - Feeds the Evaluator
    - Propagates constants of current option
- **Propagation mode**
  - Node evaluator
    - Inactive
  - Constant Propagator
    - Propagates constants of next option

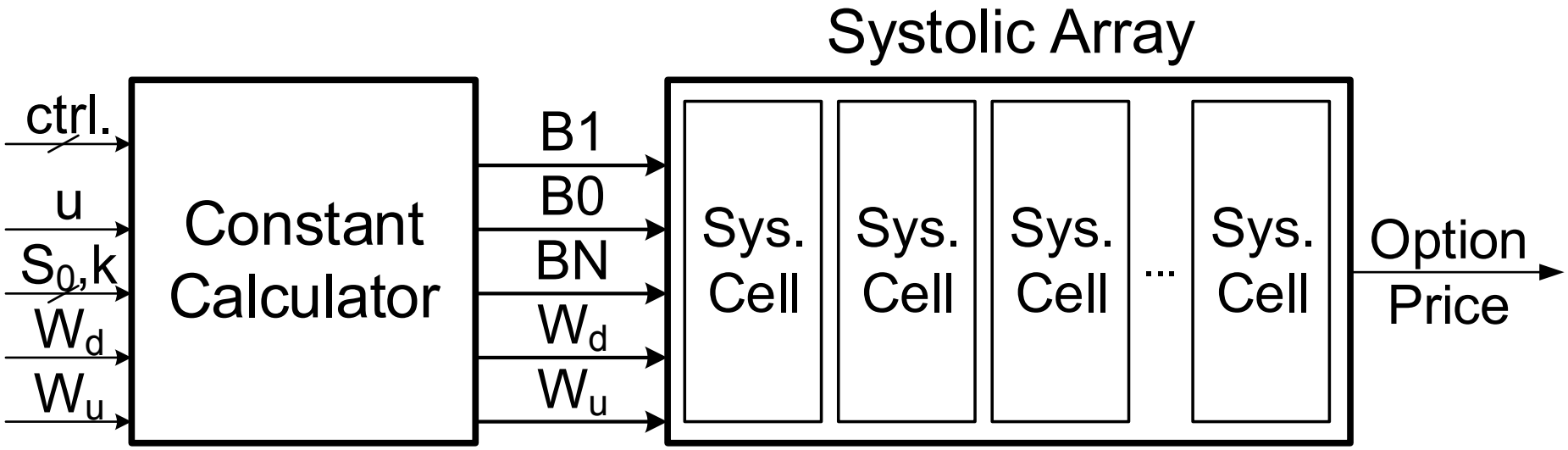


Systolic Cell: Modes of Operation

- Evaluation mode
  - Node Evaluator:
    - Active
  - Constant Propagator:
    - Feeds the Evaluator
    - Propagates constants of current option
- Propagation mode
  - Node evaluator
    - Inactive
  - Constant Propagator
    - Propagates constants of next option

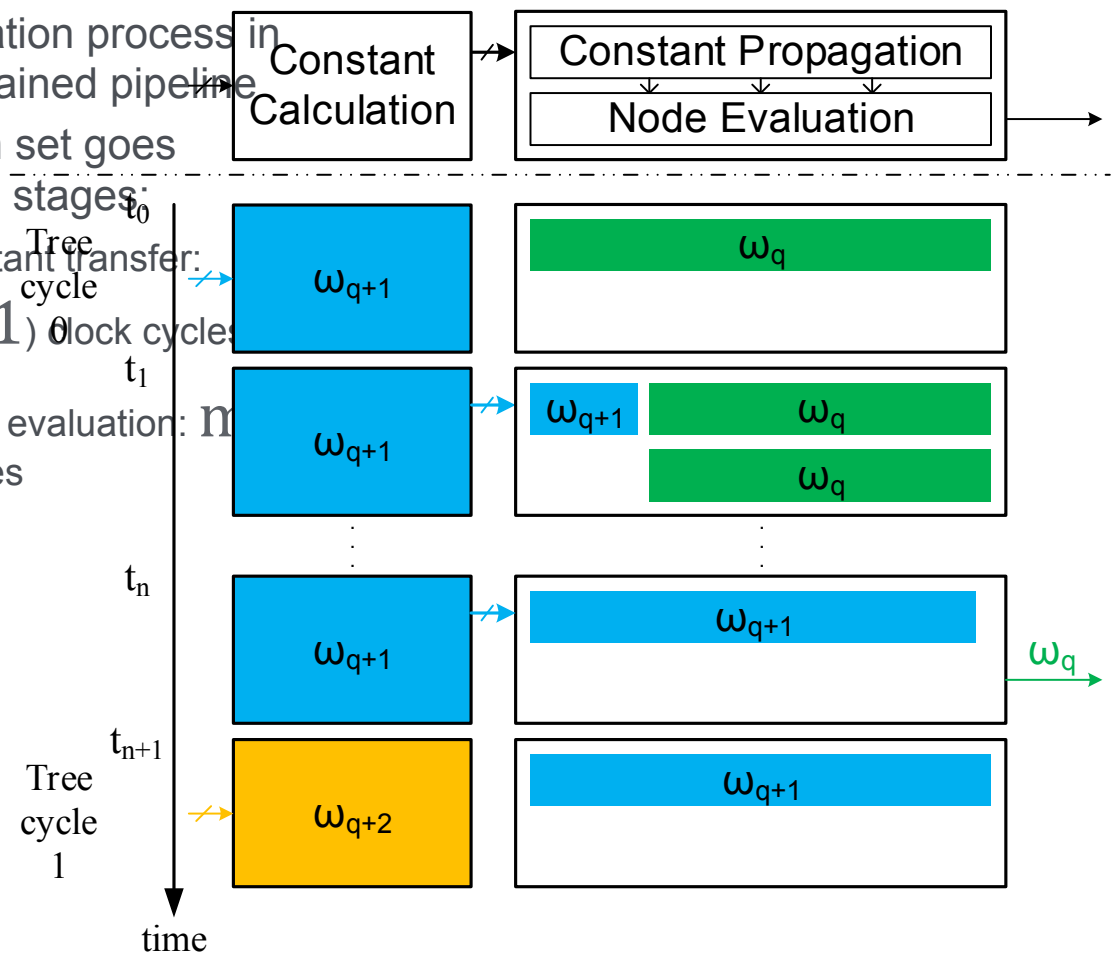


Reconfigurable System



Node Evaluation Tree Cycle

- Node evaluation process in a coarse-grained pipeline
- Every option set goes through two stages:
  - Array constant transfer: $m(n+1)$ clock cycles
  - Array node evaluation: m clock cycles



Performance Formulation

Option latency

$$L=(L\downarrow CC+m(2n+1))\times(1/f)$$

Option throughput

$$Thp=f/n+1$$

Scope

- Binomial-tree option pricing model
- Hardware architecture
- Implementation
- Results

Implementation

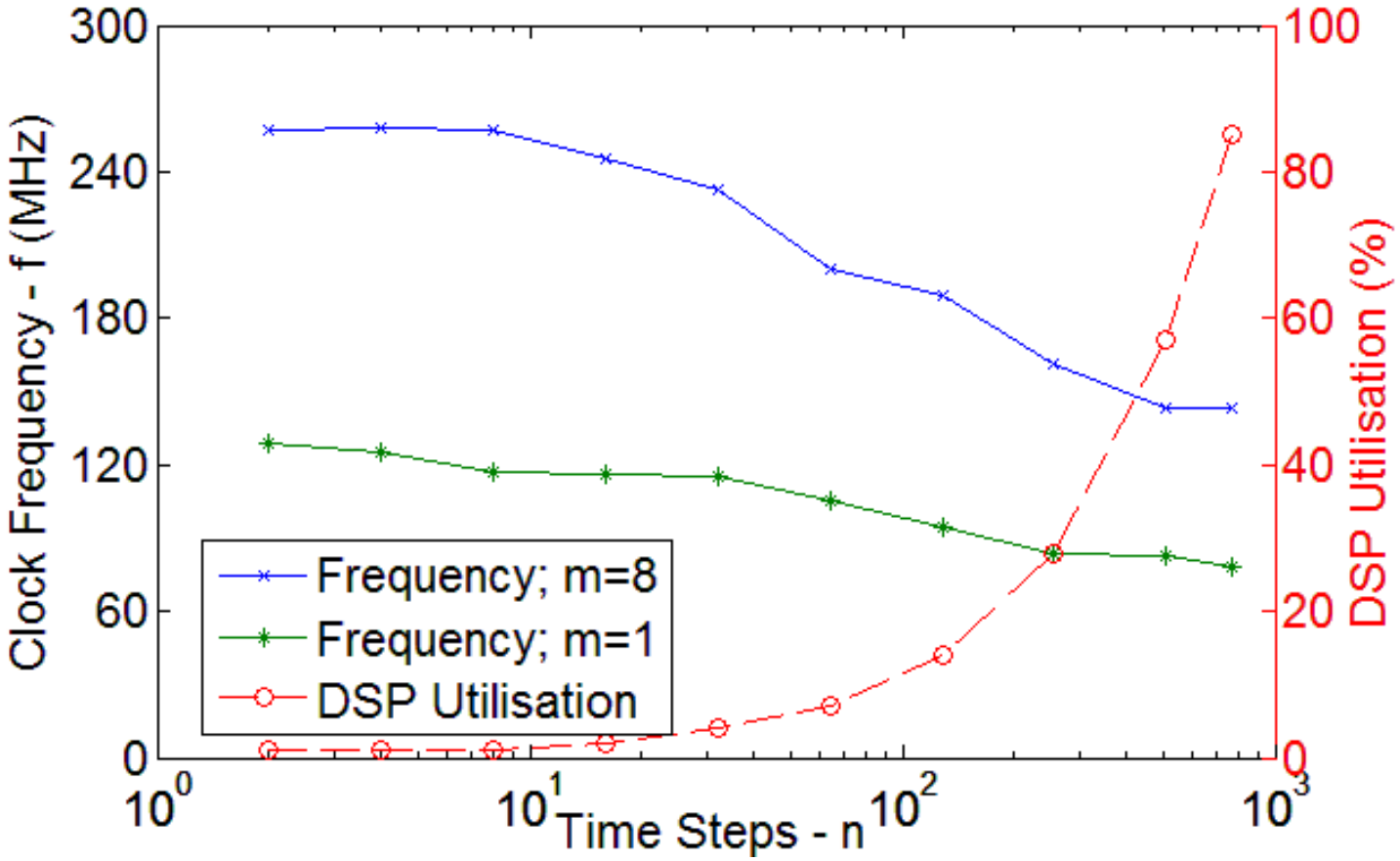
- FPGA implementation degrees of parameterisation:
  - Tree size: n
  - Concurrency configuration: m
  - Functionality of the Node Evaluator
  - Data representation
    - Fixed-point
- Design described in VHDL
- Targeted a particular accuracy of practical use

Scope

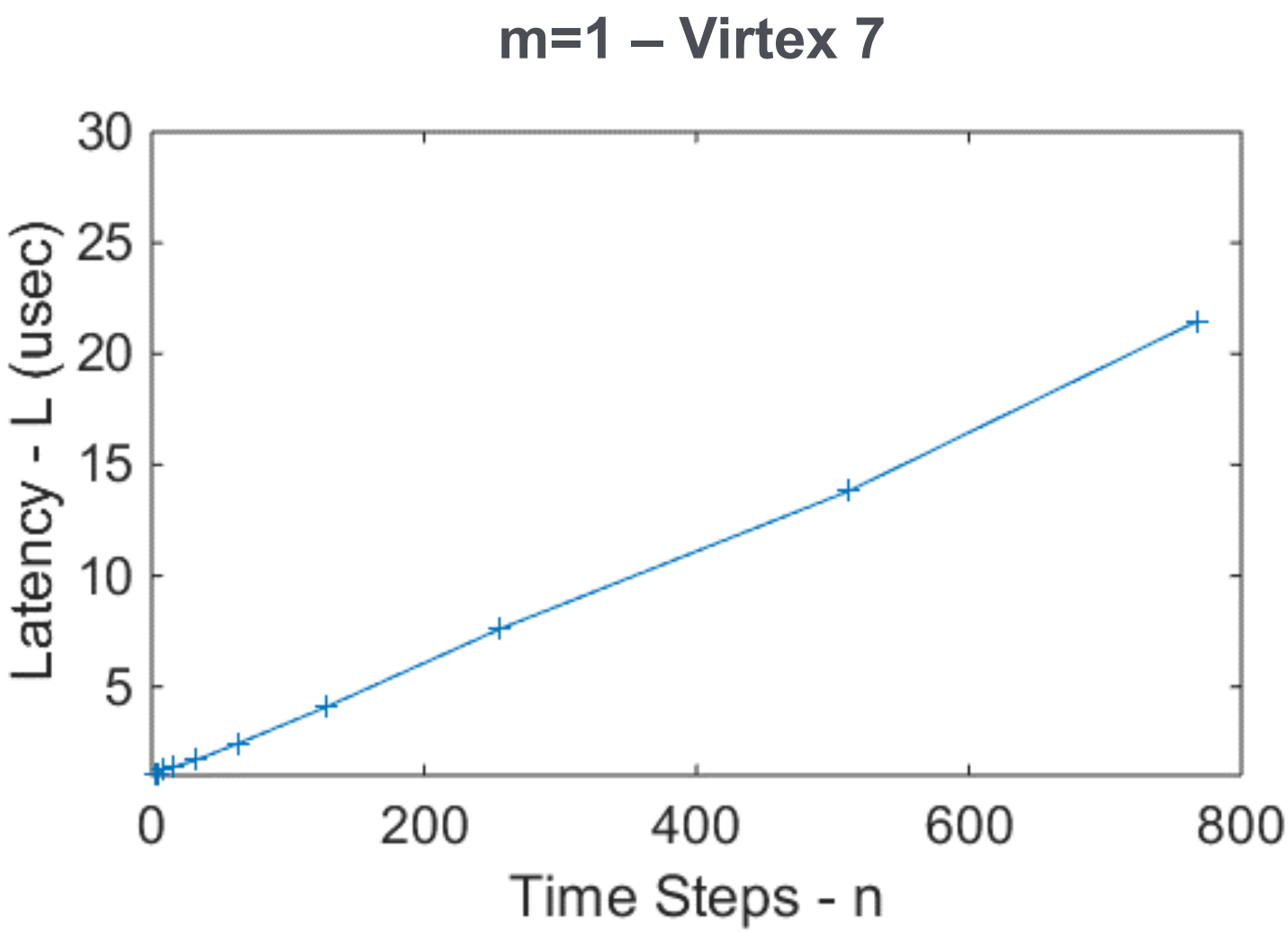
- Binomial-tree option pricing model
- Hardware architecture
- Implementation
- Results
    - Frequency-resource analysis
    - Latency figure
    - Comparison with previous approaches

Frequency-Resource Analysis

Virtex 7



Latency Figure: Micro-seconds



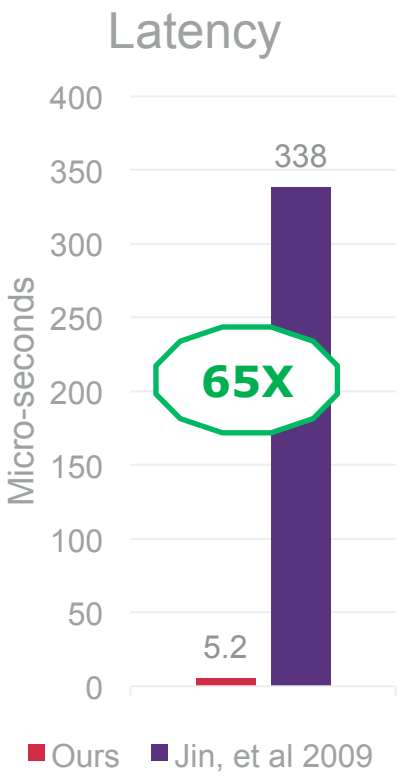
Related Works

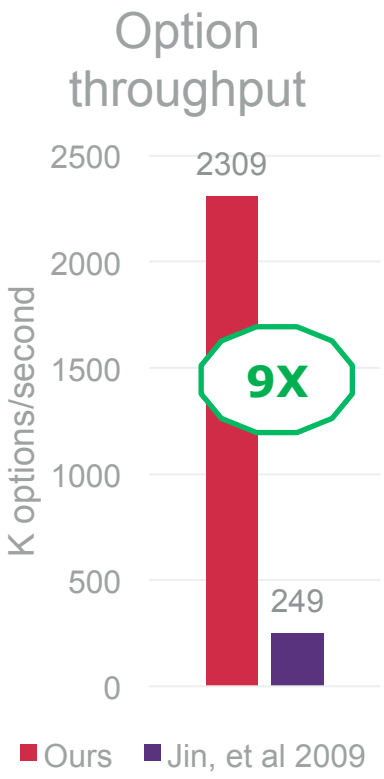
		Scalar Approach	Vector Approach	Systolic Approach
Implementation		Jin, et al 2008, 2009	Wynnyk, et al 2009; Morales, et al 2014	Ours
Resource	Computational	$O(1)$	$O(N\downarrow V)$	$O(n)$
	Memory	$O(C.n)$	$O(n)$	$O(1)$
Option latency		$O(C.n\uparrow 2)$	$O(n\uparrow 2/N\downarrow V)$	$O(n)$

Comparison: against Scalar Approach

n=96 – Fix 16.16 – Virtex 4

	Ours	Jin, et al 2009
f (MHz)	224	83
Slices	100%	99%
DSPs	78%	22%
BRAMs	3%	79%





Conclusions

- Modern large FPGAs are sufficiently large to provide spatial architectures for low-latency binomial-tree option pricing model
  - 22 micro-seconds for a 768-step tree, with a pricing rate of 100 K options/sec
- The systolic spatial architecture improved the latency figure
  - from previous quadratic methods to a linear relation
  - 65X speedup over a scalar approach
- Compile-time pipeline configuration
  - allows for higher option throughput
    - 9X improvement over a scalar approach
    - while keeping low latency
  - leaves decision on the user requirements
- More acceleration is achieved as FPGAs get bigger

Imperial College London

Low-latency option pricing using systolic binomial trees

Aryan Tavakkoli and David Thomas

Thank You!