LVars:
Lattice-based Data Structures
for Deterministic Parallel
and Distributed Programming

Lindsey Kuper
Indiana University

Intel Labs · Santa Clara, CA
March 21, 2014
Parallel systems

Distributed systems
Deterministic Parallel Programming
(observably)
Deterministic Parallel Programming
<table>
<thead>
<tr>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Book</td>
</tr>
<tr>
<td>Shoes</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>
data Item = Book | Shoes | ...

p :: IO (Map Item Int)
data Item = Book | Shoes | ...

p :: IO (Map Item Int)
p = do cart <- newIORef empty
data Item = Book | Shoes | ...

p :: IO (Map Item Int)

p = do cart <- newIORef empty
\textbf{data} Item = Book | Shoes | ...

\textbf{p} :: IO (Map Item Int)
\textbf{p} = \textbf{do} cart <- newIORef empty
data Item = Book | Shoes | ...

p :: IO (Map Item Int)
p = do cart <- newIORef empty
     async (atomicModifyIORef cart
          (\m -> (insert Book 1 m, ()))))
data Item = Book | Shoes | ...

p :: IO (Map Item Int)
p = do cart <- newIORef empty
    async (atomicModifyIORef cart
        (\m -> (insert Book 1 m, ()))))
data Item = Book | Shoes | ...

p :: IO (Map Item Int)
p = do cart <- newIORef empty
    async (atomicModifyIORef cart
        (\m -> (insert Book 1 m, ()))))
data Item = Book | Shoes | ...

p :: IO (Map Item Int)
p = do cart <- newIORef empty
async (atomicModifyIORef cart
  (\m -> (insert Book 1 m, ())))
async (atomicModifyIORef cart
  (\m -> (insert Shoes 1 m, ()))))
data Item = Book | Shoes | ...

p :: IO (Map Item Int)
p = do cart <- newIORef empty
    async (atomicModifyIORef cart
        (\m -> (insert Book 1 m, ())))
    async (atomicModifyIORef cart
        (\m -> (insert Shoes 1 m, ()))))
    res <- async (readIORef cart)
data Item = Book | Shoes | ...

p :: IO (Map Item Int)

p = do cart <- newIORef empty
    async (atomicModifyIORef cart
        (\m -> (insert Book 1 m, ()))))
    async (atomicModifyIORef cart
        (\m -> (insert Shoes 1 m, ()))))
    res <- async (readIORef cart)
    wait res
landin:lvar-examples lkuper$ make map-io-ref-data-race
ghc -O2 map-io-ref-data-race.hs -rtsopts -threaded
Linking map-io-ref-data-race ...
while true; do ./map-io-ref-data-race +RTS -N2; done
landin:lvar-examples lkuper$ make map-ioref-data-race
ghc -O2 map-ioref-data-race.hs -rtsopts -threaded
[1 of 1] Compiling Main
    (map-ioref-data-race.hs, map-ioref-data-race.o)
Linking map-ioref-data-race ...
while true; do /map-ioref-data-race +RTS -N2; done

[(Book,1),(Shoes,1)]
data Item = Book | Shoes | ...

p :: IO (Map Item Int)
p = do cart <- newIORef empty
    async (atomicModifyIORef cart
        (\m -> (insert Book 1 m, ())))
    async (atomicModifyIORef cart
        (\m -> (insert Shoes 1 m, ())))
    res <- async (readIORef cart)
    wait res
data Item = Book | Shoes | ...

p :: IO (Map Item Int)
p = do cart <- newIORef empty
    async (atomicModifyIORef cart
        (\m -> (insert Book 1 m, ())))
    async (atomicModifyIORef cart
        (\m -> (insert Shoes 1 m, ()))))
    res <- async (readIORef cart)
    wait res
data Item = Book | Shoes | ...

p :: IO (Map Item Int)
p = do cart <- newIORef empty
  a1 <- async (atomicModifyIORef cart
               (\m -> (insert Book 1 m, ())))
  async (atomicModifyIORef cart
         (\m -> (insert Shoes 1 m, ()))))
  res <- async (readIORef cart)
  wait res
data Item = Book | Shoes | ...

p :: IO (Map Item Int)

p = do cart <- newIORef empty
     a1 <- async (atomicModifyIORef cart
                    (\m -> (insert Book 1 m, ())))
     a2 <- async (atomicModifyIORef cart
                    (\m -> (insert Shoes 1 m, ()))))
     res <- async (readIORef cart)
     wait res
data Item = Book | Shoes | ... 

p :: IO (Map Item Int)

p = do cart <- newIORef empty
    a1 <- async (atomicModifyIORef cart
        (\m -> (insert Book 1 m, ())))
    a2 <- async (atomicModifyIORef cart
        (\m -> (insert Shoes 1 m, ()))))
    res <- async (do waitBoth a1 a2
      wait res     readIORef cart)
data Item = Book | Shoes | ...

p :: IO (Map Item Int)
p = do cart <- newIORef empty
    a1 <- async (atomicModifyIORef cart
      (\m -> (insert Book 1 m, ())))
ap2 <- async (atomicModifyIORef cart
      (\m -> (insert Shoes 1 m, ()))))
res <- async (do waitBoth a1 a2
      readIORef cart)
wait res
\[
\begin{align*}
p &:: \ IO \ (\text{Map} \ \text{Item} \ \text{Int}) \\
p &= \ \text{do} \\
&\quad \text{cart} \gets \ \text{newIORef} \ \text{empty} \\
&\quad a1 \gets \ \text{async} \ \text{(atomicModifyIORef} \ \text{cart} \\
&\quad \quad \quad \text{(|m|} \rightarrow \text{insert} \ \text{Book} \ 1 \ \text{m}, ())))) \\
&\quad a2 \gets \ \text{async} \ \text{(atomicModifyIORef} \ \text{cart} \\
&\quad \quad \quad \text{|m|} \rightarrow \text{insert} \ \text{Shoes} \ 1 \ \text{m}, ())) \\
&\quad \text{res} \gets \ \text{async} \ \text{(do} \ \text{waitBoth} \ a1 \ a2 \\
&\quad \quad \quad \quad \text{readIORef} \ \text{cart}) \\
&\quad \text{wait} \ \text{res} \\
&\text{main} = \ \text{do} \ v \gets p \\
&\quad \text{print} \ v
\end{align*}
\]

deterministic
p :: IO (Map Item Int)
p = do
  cart <- newIORef empty
  a1 <- async (atomicModifyIORef cart (\m -> (insert Book 1 m, ()))))
  a2 <- async (atomicModifyIORef cart (\m -> (insert Shoes 1 m, ()))))
  res <- async (do waitBoth a1 a2
                  readIORef cart)
  wait res

main = do v <- p
         print v

deterministic...now
p :: IO (Map Item Int)
p = do
  cart <- newIORef empty
  a1 <- async (atomicModifyIORef cart (\m -> (insert Book 1 m, ()))))
  a2 <- async (atomicModifyIORef cart (\m -> (insert Shoes 1 m, ()))))
  res <- async (do waitBoth a1 a2
                  readIORef cart)
  wait res

main = do v <- p
          print v

deterministic...now...we hope
\[
p ::= \text{IO (Map Item Int)}
\]
\[
p = \text{do}
\begin{align*}
\text{cart} & \leftarrow \text{newIORef empty} \\
\text{a1} & \leftarrow \text{async (atomicModifyIORef cart (m -\rightarrow (insert \text{Book 1 m, ()))))} \\
\text{a2} & \leftarrow \text{async (atomicModifyIORef cart (m -\rightarrow (insert \text{Shoes 1 m, ()})))} \\
\text{res} & \leftarrow \text{async (do waitBoth a1 a2 readIORef cart)}
\end{align*}
\]
\[
\text{main = do v} \leftarrow p \\
\text{print v}
\]

\[
p ::= \text{Par Det (IMap Item Int)}
\]
\[
p = \text{do}
\begin{align*}
\text{cart} & \leftarrow \text{newEmptyMap} \\
\text{fork (insert Book 1 cart)} \\
\text{fork (insert Shoes 1 cart)} \\
\text{return cart}
\end{align*}
\]
\[
\text{main} = \text{print (runParThenFreeze p)}
\]

\text{deterministic by construction}

[FHPC '13, POPL '14]

deterministic...now...we hope
The deterministic by construction parallel programming landscape:
The deterministic by construction parallel programming landscape:
The deterministic by construction parallel programming landscape:

- Kahn process networks
- Single-assignment
- Imperative disjoint
- \(\lambda\)-calculus

\[ f(g(x)) \triangleleft (h(y)) \]
The deterministic by construction parallel programming landscape:

- Kahn process networks
- single-assignment
- imperative disjoint
- \( \lambda \)-calculus

\[
\lambda \text{-calculus} \left< \begin{array}{c}
f (g \ x) \\
       \ &=
       \\
\end{array} \right> \]

\[
\begin{array}{c}
\text{single-assignment} \\
\text{imperative disjoint} \\
\text{Kahn process networks}
\end{array}
\]
The deterministic by construction parallel programming landscape:

Kahn process networks

single-assignment

imperative disjoint

$\lambda$-calculus

$(&\ array\ langs, ...)$

$f (g\ x) (h\ y)$
The deterministic by construction parallel programming landscape:

Kahn process networks

Kahn process networks

single-assignment

imperative disjoint

λ-calculus

λ-calculus

\( f (g x) (h y) \)

\( \text{(& array langs, ...) } \)
The deterministic by construction parallel programming landscape:

- Kahn process networks
- single-assignment
- imperative disjoint
- λ-calculus
- \( f (g \ x) \)
- \( (h \ y) \)
- (& array langs, ...)

\[ f (g \ x) \]
The deterministic by construction parallel programming landscape:

Kahn process networks

single-assignment

imperative disjoint

\( \lambda \)-calculus

\((g\ x)\ (h\ y)\)
The deterministic by construction parallel programming landscape:
The deterministic by construction parallel programming landscape:

Kahn process networks

single-assignment (IVars, CnC, ...)

imperative disjoint

\[ \lambda \text{-calculus} \]

(\& array langs, ...)

\[ f (g \ x) \]

(\& h y)

\[ g(\text{left}) \]

\[ h(\text{right}) \]
The deterministic by construction parallel programming landscape:

- Kahn process networks
- Single-assignment (IVars, CnC, ...)
- Imperative disjoint
- λ-calculus (& array langs, ...)

\[ f(g(x), h(y)) \]
The deterministic by construction parallel programming landscape:
The deterministic by construction parallel programming landscape:
The deterministic by construction parallel programming landscape:

Can we generalize and unify these points on the map?
The deterministic by construction parallel programming landscape:

- Kahn process networks
- Single-assignment (IVars, CnC, ...)
- Imperative disjoint (DPJ, ...)
- λ-calculus (array langs, ...)

Can we generalize and unify these points on the map? Yes!
data Item = Book | Shoes | ...

p :: IO (Map Item Int)
p = do cart <- newIORef empty
    async (atomicModifyIORef cart
        (\m -> (insert Book 1 m, ())))
    async (atomicModifyIORef cart
        (\m -> (insert Shoes 1 m, ()))))
    res <- async (readIORef cart)
    wait res
data Item = Book | Shoes | ...

p :: IO (Map Item Int)
p = do cart <- newIORef empty
    async (atomicModifyIORef cart
    (\m -> (insert Book 1 m, ())))
    async (atomicModifyIORef cart
    (\m -> (insert Shoes 1 m, ())))
    res <- async (readIORef cart)
    wait res
\textbf{data} Item = Book | Shoes | ...

\textbf{p} :: IO (Map Item Int)
\textbf{p} = do cart <- newIORef empty
async (atomicModifyIORef cart
  (\_m -> (insert \textbf{Book} 1 m, ())))
async (atomicModifyIORef cart
  (\_m -> (insert \textbf{Shoes} 1 m, ())))
res <- async (readIORef cart)
wait res

\textbf{IVars: single writes, blocking (but exact) reads}

[Arvind et al., 1989]
\textbf{data} Item = Book | Shoes | ... \\

\textbf{p} :: IO (Map Item Int) \\
\textbf{p} = do cart <- newIORef empty \\
    async (atomicModifyIORef cart \\
        \( \text{\textbackslash m \rightarrow (insert \hspace{0.5em} \text{Book} \hspace{0.5em} 1 \hspace{0.5em} m, ())}) \)) \\
    async (atomicModifyIORef cart \\
        \( \text{\textbackslash m \rightarrow (insert \hspace{0.5em} \text{Shoes} \hspace{0.5em} 1 \hspace{0.5em} m, ())}) \)) \\
    res <- async (readIORef cart) \\
    wait res \\

IVars: single writes, blocking (but exact) reads \\
[Arvind et al., 1989]
data Item = Book | Shoes | ...

p :: IO (Map Item Int)
p = do  cart <- newIORef empty
           async (atomicModifyIORef cart (\m -> (insert Book 1 m, ())))
           async (atomicModifyIORef cart (\m -> (insert Shoes 1 m, ()))))
res <- async (readIORef cart)
wait res

IVars: single writes, blocking (but exact) reads
[Arvind et al., 1989]

LVars: multiple least-upper-bound writes, blocking threshold reads
[FHPC ’13]
data Item = Book | Shoes | ...

p :: IO (Map Item Int)

p = do cart <- newIORef empty
    async (atomicModifyIORef cart
      (\m -> (insert Book 1 m, ())))
    async (atomicModifyIORef cart
      (\m -> (insert Shoes 1 m, ()))))
    res <- async (readIORef cart)
    wait res

IVars: single writes, blocking (but exact) reads
[Arvind et al., 1989]

LVars: multiple least-upper-bound writes, blocking threshold reads
[FHPC ’13]

* actually a bounded join-semilattice
```plaintext

Raises an error, since $3 \sqcup 4 = \top$

```do

fork (put num 3)

fork (put num 4)

```

Works fine, since $4 \sqcup 4 = 4$

```do

fork (put num 4)

fork (put num 4)

```
data Item = Book | Shoes | ...

p = do
  cart <- newEmptyMap
  fork (insert Shoes 1 cart)
  fork (insert Book 2 cart)
  getKey Book cart -- returns 2
\[
data \text{ Item} = \text{ Book} \mid \text{ Shoes} \mid \ldots
\]

\[
p = \text{ do}
\begin{align*}
\text{cart } &\leftarrow \text{ newEmptyMap} \\
\text{fork (insert Shoes 1 cart)} \\
\text{fork (insert Book 2 cart)} \\
\text{getKey Book cart} \quad \text{-- returns } 2
\end{align*}
\]
data Item = Book | Shoes | ...

p = do
    cart <- newEmptyMap
    fork (insert Shoes 1 cart)
    fork (insert Book 2 cart)
    getKey Book cart -- returns 2
data Item = Book | Shoes | ...

p = do
  cart <- newEmptyMap
  fork (insert Shoes 1 cart)
  fork (insert Book 2 cart)
  getKey Book cart -- returns 2
data Item = Book | Shoes | ...

p = do
    cart <- newEmptyMap
    fork (insert Shoes 1 cart)
    fork (insert Book 2 cart)
    getKey Book cart -- returns 2
data Item = Book | Shoes | ...

p = do
   cart <- newEmptyMap
   fork (insert Shoes 1 cart)
   fork (insert Book 2 cart)
   getKey Book cart -- returns 2
data Item = Book | Shoes | ...

p = do
cart <- newEmptyMap
fork (insert Shoes 1 cart)
fork (insert Book 2 cart)
getKey Book cart -- returns 2
data Item = Book | Shoes | ...

p = do
  cart <- newEmptyMap
  fork (insert Shoes 1 cart)
  fork (insert Book 2 cart)
  getKey Book cart -- returns 2
data Item = Book | Shoes | ...

\{(Book, 1), (Book, 2), \ldots\}

p = do
    cart <- newEmptyMap
    fork (insert Shoes 1 cart)
    fork (insert Book 2 cart)
    getKey Book cart -- returns 2
data Item = Book | Shoes | ...

\{(Book, 1), (Book, 2), \ldots\}

p = do
  cart <- newEmptyMap
  fork (insert Shoes 1 cart)
  fork (insert Book 2 cart)
  getKey Book cart -- returns 2
The threshold set must be pairwise incompatible.
The threshold set must be pairwise incompatible.

The data set is:
{(Book, 1), (Book, 2), ...}

The proof obligation is:

```plaintext
data Item = Book | Shoes | ...
p = do
cart <- newEmptyMap
fork (insert Shoes 1 cart)
fork (insert Book 2 cart)
getKey Book cart -- returns 2
```
The threshold set must be pairwise incompatible.

```
{ (Book, 1), (Book, 2), ... }
```

Data Item = Book | Shoes | ...

p = do
  cart <- newEmptyMap
  fork (insert Shoes 1 cart)
  fork (insert Book 2 cart)
  getKey Book cart -- returns 2
seen nodes
seen nodes

0  1  3
4  5  6
7  9  10
11

already seen

0  1  2  3
4  5  6  7
8  9  10  11
seen nodes

0  1  3
4  5  6
7  9  10
11

already seen

0
1
2
already seen

seen nodes

0 1 3
4 5 6
7 9 10
11

...
Events are updates that change an LVar’s state
Events are updates that change an LVar's state
Event handlers listen for events and launch callbacks in response
Events are updates that change an LVar's state

Event handlers listen for events and launch callbacks in response

```
traverse g startNode = do
```
Events are updates that change an LVar's state
Event handlers listen for events and launch callbacks in response

```
traverse g startNode = do
    seen <- newEmptySet
```
Events are updates that change an LVar's state.

Event handlers listen for events and launch callbacks in response.

```haskell
traverse g startNode = do
  seen <- newEmptySet
  h <- newHandler seen
      (\node -> do
          mapM (\v -> insert v seen) (neighbors g node)
          return ())
```
Events are updates that change an LVar's state

Event handlers listen for events and launch callbacks in response

```haskell
traverse g startNode = do
    seen <- newEmptySet
    h <- newHandler seen
        (\node -> do
            mapM (\v -> insert v seen) (neighbors g node)
            return ()
        )
    insert startNode seen
```
Events are updates that change an LVar's state
Event handlers listen for events and launch callbacks in response
**quiesce** blocks until all callbacks launched by a given handler are done running

```
traverse g startNode = do
  seen <- newEmptySet
  h <- newHandler seen
    (\node -> do
      mapM (\v -> insert v seen)
        (neighbors g node)
      return ()
    )
  insert startNode seen
```
**Events** are updates that change an LVar's state

**Event handlers** listen for events and launch callbacks in response.

**quiesce** blocks until all callbacks launched by a given handler are done running.

```
traverse g startNode = do
  seen <- newEmptySet
  h <- newHandler seen
      (\node -> do
        mapM (\v -> insert v seen)
            (neighbors g node)
        return ()
      )
  insert startNode seen
  quiesce h
```
Events are updates that change an LVar's state

Event handlers listen for events and launch callbacks in response

**quiesce** blocks until all callbacks launched by a given handler are done running

---

```haskell
traverse g startNode = do
  seen <- newEmptySet
  h <- newHandler seen
  (\node -> do
    mapM (\v -> insert v seen)
    (neighbors g node)
    return ()
  )
  insert startNode seen
  quiesce h
...
```
**freeze**: exact non-blocking read

```haskell
traverse g startNode = do
  seen <- newEmptySet
  h <- newHandler seen
  (\node -> do
    mapM (\v -> insert v seen) (neighbors g node)
    return ()
  )
  insert startNode seen
  quiesce h
...
```
freeze: exact non-blocking read
Attempts to write to a frozen LVar raise a write-after-\texttt{freeze} exception

\begin{verbatim}
traverse g startNode = do
  seen <- newEmptySet
  h <- newHandler seen
  (\node -> do
    mapM (\v -> insert v seen)
    (neighbors g node)
    return ())
  insert startNode seen
  quiesce h
...
\end{verbatim}
**freeze**: exact non-blocking read

Attempts to write to a frozen LVar raise a write-after-**freeze** exception

```
traverse g startNode = do
  seen <- newEmptySet
  h <- newHandler seen
  (\node -> do
     mapM (\v -> insert v seen)
         (neighbors g node)
     return ())
  insert startNode seen
  quiesce h
  freeze seen
```
freeze: exact non-blocking read
Attempts to write to a frozen LVar raise a write-after-freeze exception
Two possible outcomes: either the same final value or an exception

```haskell
traverse g startNode = do
  seen <- newEmptySet
  h <- newHandler seen
  (\node -> do
    mapM (\v -> insert v seen)
      (neighbors g node)
    return ()
  )
insert startNode seen
quiesce h
freeze seen
```
freeze: exact non-blocking read
Attempts to write to a frozen LVar raise a write-after-freeze exception
Two possible outcomes: either the same final value or an exception

Theorem 1 (Quasi-Determinism). If $\sigma \xrightarrow{*} \sigma'$ and $\sigma \xrightarrow{*} \sigma''$, and neither $\sigma'$ nor $\sigma''$ can take a step, then either:
1. $\sigma' = \sigma''$ up to a permutation on locations $\pi$, or
2. $\sigma' = \text{error}$ or $\sigma'' = \text{error}$.

[POPL '14]
**freeze**: exact non-blocking read

Attempts to write to a frozen LVar raise a write-after-**freeze** exception

Two possible outcomes: either the same final value or an exception

---

**Theorem 1** (Quasi-Determinism). If \( \sigma \xrightarrow{*} \sigma' \) and \( \sigma \xrightarrow{*} \sigma'' \), and neither \( \sigma' \) nor \( \sigma'' \) can take a step, then either:

1. \( \sigma' = \sigma'' \) up to a permutation on locations \( \pi \), or
2. \( \sigma' = \text{error} \) or \( \sigma'' = \text{error} \). 

[POPL '14]
freeze: exact non-blocking read
Attempts to write to a frozen LVar raise a write-after-f\texttt{freeze} exception
Two possible outcomes: either the same final value or an exception

\textbf{Theorem 1} (Quasi-Determinism). If $\sigma \xrightarrow{} \sigma'$ and $\sigma \xrightarrow{} \sigma''$, and neither $\sigma'$ nor $\sigma''$ can take a step, then either:

1. $\sigma' = \sigma''$ up to a permutation on locations $\pi$, or
2. $\sigma' = \text{error}$ or $\sigma'' = \text{error}$.

[POPL '14]
freeze: exact non-blocking read
Attempts to write to a frozen LVar raise a write-after-freeze exception
Two possible outcomes: either the same final value or an exception

Theorem 1 (Quasi-Determinism). If \( \sigma \xrightarrow{*} \sigma' \) and \( \sigma \xrightarrow{*} \sigma'' \), and neither \( \sigma' \) nor \( \sigma'' \) can take a step, then either:
1. \( \sigma' = \sigma'' \) up to a permutation on locations \( \pi \), or
2. \( \sigma' = \text{error} \) or \( \sigma'' = \text{error} \).

```haskell
(define-LVish-language nat downset-op max natural)
```

```
insert v seen)
insert startNode seen
quiesce h
freeze seen
```

[POPL '14]
**freeze**: exact non-blocking read

Attempts to write to a frozen LVar raise a write-after-**freeze** exception

Two possible outcomes: either the same final value or an exception

---

**Theorem 1** (Quasi-Determinism). *If* $\sigma \xrightarrow{\ast} \sigma'$ *and* $\sigma \xrightarrow{\ast} \sigma''$, *and neither* $\sigma'$ *nor* $\sigma''$ *can take a step, then either:*

1. $\sigma' = \sigma''$ *up to a permutation on locations* $\pi$, *or*
2. $\sigma' = \text{error}$ *or* $\sigma'' = \text{error}$.

---

```
insert v seen)
(insert startNode seen)
return ()
```

or error:

```
[(Book,1), (Shoes,1)]
```

or error:

```
[(Bo,1)]
```

or error:

```
[(Sh,1)]
```

---

*[POPL ’14]*
**freeze**: exact non-blocking read

Attempts to write to a frozen LVar raise a write-after-**freeze** exception

Two possible outcomes: either the same final value or an exception

```haskell
traverse g startNode = do
  seen <- newEmptySet
  h <- newHandler seen
  (\node -> do
    mapM (\v -> insert v seen)
    (neighbors g node)
    return ()
  )
  insert startNode seen
  quiesce h
  freeze seen
```
**freeze**: exact non-blocking read

Attempts to write to a frozen LVar raise a write-after-**freeze** exception

Two possible outcomes: either the same final value or an exception

```haskell
traverse g startNode = do
  seen <- newEmptySet
  h <- newHandler seen
  (\node -> do
    mapM (\v -> insert v seen)
    (neighbors g node)
    return ()
  )
  insert startNode seen
  quiesce h
  freeze seen
```
**freeze**: exact non-blocking read

Attempts to write to a frozen LVar raise a write-after-**freeze** exception

Two possible outcomes: either the same final value or an exception

```haskell
traverse g startNode = do
  seen <- newEmptySet
  h <- newHandler seen
  (\node -> do
    mapM (\v -> insert v seen)
    (neighbors g node)
    return ()
  )
  insert startNode seen
  quiesce h
  freeze seen
```

Let the system handle this for us:
**freeze**: exact non-blocking read

Attempts to write to a frozen LVar raise a write-after-`freeze` exception

Two possible outcomes: either the same final value or an exception

```haskell
traverse g startNode = do
  seen <- newEmptySet
  h <- newHandler seen
  (\node -> do
    mapM (\v -> insert v seen)
    (neighbors g node)
    return ()
  )
  insert startNode seen
  quiesce h
  freeze seen
```

Let the system handle this for us:

`runParThenFreeze`
LVish

a Haskell library for parallel programming with LVars
LVish

a Haskell library for parallel programming with LVars

LVar operations run in Par computations
LVish

a Haskell library for parallel programming with LVars

LVar operations run in \texttt{Par} computations

Lightweight threads
LVish

a Haskell library for parallel programming with LVars

LVar operations run in \texttt{Par} computations

Lightweight threads

\texttt{Par} computations indexed by effect level

\begin{verbatim}
p :: Par Det (IMap Item Int)
p = do
cart <- newEmptyMap
fork (insert Book 1 cart)
fork (insert Shoes 1 cart)
return cart
\end{verbatim}
LVish
a Haskell library for parallel programming with LVars

LVars operations run in Par computations

Lightweight threads

Par computations indexed by effect level

runParThenFreeze captures the freeze-after-writing idiom

```haskell
p :: Par Det (IMap Item Int)
p = do
  cart <- newEmptyMap
  fork (insert Book 1 cart)
  fork (insert Shoes 1 cart)
  return cart
```
LVish

a Haskell library for parallel programming with LVars

LVar operations run in \texttt{Par} computations

Lightweight threads

\texttt{Par} computations indexed by \texttt{effect level}

\texttt{runParThenFreeze} captures the freeze-after-writing idiom

\begin{verbatim}
p :: Par Det (IMap Item Int)
p = do
cart <- newEmptyMap
fork (insert Book 1 cart)
fork (insert Shoes 1 cart)
return cart

main = print (runParThenFreeze p)
\end{verbatim}
LVish
a Haskell library for parallel programming with LVars

LVar operations run in Par computations
Lightweight threads
Par computations indexed by effect level
runParThenFreeze captures the freeze-after-writing idiom
Efficient lock-free sets, maps, etc.

```haskell
p :: Par Det (IMap Item Int)
p = do
  cart <- newEmptyMap
  fork (insert Book 1 cart)
  fork (insert Shoes 1 cart)
  return cart

main = print (runParThenFreeze p)
```
LVish
a Haskell library for parallel programming with LVars

LVar operations run in Par computations
Lightweight threads
Par computations indexed by effect level
runParThenFreeze captures the freeze-after-writing idiom
Efficient lock-free sets, maps, etc.
Implement your own LVars, too

```
p :: Par Det (IMap Item Int)
p = do
cart <- newEmptyMap
fork (insert Book 1 cart)
fork (insert Shoes 1 cart)
return cart

main = print (runParThenFreeze p)
```
LVish
a Haskell library for parallel programming with LVars

LVar operations run in **Par** computations
Lightweight threads
**Par** computations indexed by effect level
**runParThenFreeze** captures the freeze-after-writing idiom
Efficient lock-free sets, maps, etc.
Implement your own LVars, too

```
(p :: Par Det (IMap Item Int))
p = do
  cart <- newEmptyMap
  fork (insert Book 1 cart)
  fork (insert Shoes 1 cart)
  return cart

main = print (runParThenFreeze p)
```

[https://hackage.haskell.org/package/lvish](https://hackage.haskell.org/package/lvish)
Deterministic Parallel Programming
(observably)
Deterministic Parallel Programming
Deterministic Parallel Programming

(observably)  (irregular)
Case study: $k$-CFA static analysis parallelized with LVish [POPL ’14]

[Earl et al., ICFP ’12]
Case study: *k*-CFA static analysis parallelized with LVish [POPL ’14]

- up to 20x speedup, even on one core, from not having to copy data

[Earl et al., ICFP ’12]
Case study: $k$-CFA static analysis parallelized with LVish [POPL ’14]

- up to 20x speedup, even on one core, from not having to copy data
- 7-8x parallel speedup
Case study: k-CFA static analysis parallelized with LDVish [POPL ’14]

- up to 20x speedup, even on one core, from not having to copy data
- 7-8x parallel speedup
Case study: $k$-CFA static analysis parallelized with LVish [POPL ’14]

- up to 20x speedup, even on one core, from not having to copy data
- 7-8x parallel speedup
- Lock-free structures help
Case study: $k$-CFA static analysis parallelized with LVish [POPL ’14]

- up to 20x speedup, even on one core, from not having to copy data
- 7-8x parallel speedup
- Lock-free structures help
Case study: $k$-CFA static analysis parallelized with LVish [POPL ’14]

- up to 20x speedup, even on one core, from not having to copy data
- 7-8x parallel speedup
- Lock-free structures help
Case study: $k$-CFA static analysis parallelized with LVish [POPL ’14]

- up to 20x speedup, even on one core, from not having to copy data
- 7-8x parallel speedup
- Lock-free structures help
Case study: k-CFA static analysis parallelized with LVish [POPL ’14]

- up to 20x speedup, even on one core, from not having to copy data
- 7-8x parallel speedup
- Lock-free structures help

See also: Phylogenetic tree binning parallelized with LVish [PLDI ’14]
LVars and LVish across the landscape:

Kahn process networks

G → H

single-assignment

imperative disjoint

λ-calculus

f (g x) (h y)

g(left) h(right)
LVars and LVish across the landscape:

Kahn process networks

single-assignment

imperative disjoint

λ-calculus

\[ f(g(x), h(y)) \]
Kahn process networks

single-assignment

imperative disjoint

$\lambda$-calculus

LVars and LVish across the landscape:
LVars and LVish across the landscape:

- Kahn process networks
- single-assignment
- imperative disjoint
- λ-calculus

\[ f(g(x), h(y)) \]
LVars and LVish across the landscape:

Kahn process networks

single-assignment

imperative disjoint

λ-calculus

\[ f (g \ x) \quad (h \ y) \]

\[ g(\text{left}) \quad h(\text{right}) \]
LVars and LVish across the landscape:
Parallel systems

Distributed systems
Distributed systems
Figure 1: Global Overseas Distribution of Books

The figure illustrates the distribution of books across different regions of the world. Each node represents a database or a distribution center, and the arrows indicate the flow of books from one location to another.

- **getKey** Book
- **getKey** Book
- **getKey** Book
- **getKey** Book
- **getKey** Book

This visualization highlights how books are systematically distributed across international boundaries, ensuring accessibility to readers globally.
Eventual consistency.
Eventual consistency. How?
Dynamo: Amazon’s Highly Available Key-value Store
Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall and Werner Vogels
Amazon.com

ABSTRACT
Reliability at massive scale is one of the biggest challenges we face at Amazon.com, one of the largest e-commerce operations in the world; even the slightest outage has significant financial consequences and impacts customer trust. The Amazon.com platform, which provides services for many web sites worldwide, is implemented on top of an infrastructure of tens of thousands of servers and network components located in many datacenters around the world. At this scale, small and large components fail continuously and the way persistent state is managed in the face of these failures drives the reliability and scalability of the software systems.

This paper presents the design and implementation of Dynamo, a highly available key-value storage system that some of Amazon’s core services use to provide an “always-on” experience. To achieve this level of availability, Dynamo sacrifices consistency under certain failure scenarios. It makes extensive use of object versioning and application-assisted conflict resolution in a manner that provides a novel interface for developers to use.

Categories and Subject Descriptors
D.4.2 [Operating Systems]: Storage Management; D.4.5 [Operating Systems]: Reliability; D.4.2 [Operating Systems]: Performance;

General Terms

1. INTRODUCTION
Amazon runs a world-wide e-commerce platform that serves tens of millions customers at peak times using tens of thousands of servers located in many data centers around the world. There are strict operational requirements on Amazon’s platform in terms of performance, reliability and efficiency, and to support continuous growth the platform needs to be highly scalable. Reliability is one of the most important requirements because even the slightest outage has significant financial consequences and impacts customer trust. In addition, to support continuous growth, the platform needs to be highly scalable.

One of the lessons our organization has learned from operating Amazon’s platform is that the reliability and scalability of a system is dependent on how its application state is managed. Amazon uses a highly decentralized, loosely coupled, service oriented architecture consisting of hundreds of services. In this environment there is a particular need for storage technologies that are always available. For example, customers should be able to view and add items to their shopping cart even if disks are failing, network routes are flapping, or data centers are being destroyed by tornados. Therefore, the service responsible for managing shopping carts requires that it can always write to and read from its data store, and that its data needs to be available across multiple data centers.

Dealing with failures in an infrastructure comprised of millions of components is our standard mode of operation; there are always a small but significant number of server and network components that are failing at any given time. As such Amazon’s software systems need to be constructed in a manner that treats failure handling as the normal case without impacting availability or performance.

To meet the reliability and scaling needs, Amazon has developed a number of storage technologies, of which the Amazon Simple Storage Service (also available outside of Amazon and known as Amazon S3), is probably the best known. This paper presents the design and implementation of Dynamo, another highly available and scalable distributed data store built for Amazon’s platform. Dynamo is used to manage the state of services that have very high reliability requirements and need tight control over the tradeoffs between availability, consistency, cost-effectiveness and performance. Amazon’s platform has a very diverse set of applications with different storage requirements. A select set of applications requires a storage technology that is flexible enough to let application designers configure their data store appropriately based on these tradeoffs to achieve high availability and guaranteed performance in the most cost effective manner.

There are many services on Amazon’s platform that only need primary-key access to a data store. For many services, such as those that provide best seller lists, shopping carts, customer preferences, session management, sales rank, and product catalog, the common pattern of using a relational database would lead to inefficiencies and limit scale and availability. Dynamo provides a simple primary-key only interface to meet the requirements of these applications.

Dynamo uses a synthesis of well known techniques to achieve scalability and availability: Data is partitioned and replicated using consistent hashing [10], and consistency is facilitated by object versioning [12]. The consistency among replicas during updates is maintained by a quorum-like technique and a decentralized replica synchronization protocol. Dynamo employs
since the application is aware of the data schema it can decide on the conflict resolution method that is best suited for its client’s experience. For instance, the application that maintains customer shopping carts can choose to “merge” the conflicting versions and return a single unified shopping cart.
Conflict-Free Replicated Data Types*

Marc Shapiro1,2, Nuno Preguiça1,2, Carlos Baquero3, and Marek Zawirski1,4

1 INRIA, Paris, France
2 CTTI, Universidade Nova de Lisboa, Portugal
3 Universidade do Minho, Portugal
4 UPMC, Paris, France

Abstract. Replicating data under Eventual Consistency (EC) allows any replica to accept updates without remote synchronisation. This ensures performance and scalability in large-scale distributed systems (e.g., clouds). However, published EC approaches are ad-hoc and error-prone. Under a formal Strong Eventual Consistency (SEC) model, we study sufficient conditions for convergence. A data type that satisfies those conditions is called a Conflict-free Replicated Data Type (CRDT). Replicas of any CRDT are guaranteed to converge in a self-simulating manner, despite any number of failures. This paper formalises two popular approaches (state- and operation-based) and their relevant sufficient conditions. We study a number of useful CRDTs, such as sets with clean semantics, supporting both add and remove operations, and consider in depth the more complex Graph data type. CRDT types can be composed to develop large-scale distributed applications, and have interesting theoretical properties.

Keywords: Eventual Consistency, Replicated Shared Objects, Large-Scale Distributed Systems.

1 Introduction

Replication and consistency are essential features of any large distributed system, such as the WWW, peer-to-peer, or cloud computing platforms. The standard “strong consistency” approach serialises updates in a global total order [10]. This constitutes a performance and scalability bottleneck. Furthermore, strong consistency conflicts with availability and partition-tolerance [8]. When network delays are large or partitioning is an issue, as in delay-tolerant networks, disconnected operation, cloud computing, or P2P systems, eventual consistency promises better availability and performance [17,21]. An update executes at some replica, without synchronisation; later, it is sent to the other replicas. It is guaranteed that all the replicas will eventually receive the update or be notified that the update is not made, if the protocol designed for this is adequate. A more reliable protocol would be expected for data that is not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Dynamo: Amazon’s Highly Available Key-value Store

Giuseppe DeCandia, Deniz Hasto, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Voss, and Werner Vogels

Abstract. Reliability at massive scale is one of the biggest challenges we face at Amazon.com, one of the largest e-commerce operations in the world; even the slightest outage has significant financial consequences and impacts customer trust. The Amazon.com platform, which provides services for many web sites worldwide, is implemented on top of an infrastructure of thousands of servers and network components located in many datacenters around the world. At this scale, small and large components fail continuously and the way persistent state is managed in the face of these failures drives the reliability and scalability of the software systems.

This paper presents the design and implementation of Dynamo, a highly available key-value storage system that some of Amazon’s core services use to provide an “always-on” experience. To achieve this level of availability, Dynamo sacrifices consistency under certain failure scenarios. It makes extensive use of object versioning and application-assisted conflict resolution in a manner that provides a novel interface for developers to use.

Categories and Subject Descriptors

C.2 [Computer Systems Organization]: Distributed Systems

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Additional Resources

Amazon S3: http://aws.amazon.com/s3/

Dynamo uses a synthesis of well-known techniques to achieve scalability and availability. Data is partitioned and replicated using consistent hashing [10], and consistency is facilitated by object versioning [12]. The consistency among replicas during updates is maintained by a quorum-like technique and a decentralized replica synchronization protocol. Dynamo employs
Conflict-Free Replicated Data Types*

Marc Shapiro1,5, Nuno Preguiça1,2, Carlos Baquero3, and Marek Zawirski1,4
1 INRIA, Paris, France
2 CITI, Universidade Nova de Lisboa, Portugal
3 Universidade do Minho, Portugal
4 UPMC, Paris, France
5 LIP6, Paris, France

Abstract. Replicating data under Eventual Consistency (EC) allows any replica to accept updates without remote synchronisation. This ensures performance and scalability in large-scale distributed systems (e.g., clouds). However, published EC approaches are ad-hoc and error-prone. Under a formal Strong Eventual Consistency (SEC) model, we study sufficient conditions for convergence. A data type that satisfies those conditions is called a Conflict-free Replicated Data Type (CRDT). Replicas of any CRDT are guaranteed to converge to a self-stabilising manner, despite any number of failures. This paper formalises two popular approaches (state- and operation-based) and their relevant sufficient conditions. We study a number of useful CRDTs, such as sets with clean semantics, supporting both add and remove operations, and consider in depth the more complex Graph data type. CRDT types can be composed to develop large-scale distributed applications, and have interesting theoretical properties.

Keywords: Eventual Consistency, Replicated Shared Objects, Large-Scale Distributed Systems.

1 Introduction

Replication and consistency are essential features of any large distributed system, such as the WWW, peer-to-peer, or cloud computing platforms. The standard “strong consistency” approach serialises updates in a global total order [10]. This constitutes a performance and scalability bottleneck. Furthermore, strong consistency conflicts with availability and partition-tolerance [8].

When network delays are large or partitioning is an issue, as in delay-tolerant networks, disconnected operation, cloud computing, or P2P systems, eventual consistency promises better availability and performance [17,21]. An update executes at some replica, without synchronisation; later, it is sent to the other

Dynamo: Amazon’s Highly Available Key-value Store

Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall and Werner Vogels
Amazon.com

ABSTRACT

One of the lessons our organization has learned from operating Amazon’s platform is that the reliability and scalability of a system is dependent on how its application state is managed. Amazon uses a highly decentralized, loosely coupled, service oriented architecture consisting of hundreds of services. In this environment there is a particular need for storage technologies that are always available. For example, customers should be able to view and add items to their shopping cart even if disk are failing, network routes are flapping, or data centers are being shut down by tornadoes. Therefore, the service responsible for managing shopping cart requires that it can always write to and from its data store, and that its data need to be available to multiple data centers.

Since most failures in an infrastructure comprised of millions of components are short and not significant, some number of server and network components are always available. As such, Amazon’s software must be constructed in a manner that treats failure as the normal case without impacting availability or performance.

In meeting the reliability and scaling needs, Amazon has developed Dynamo, another highly available key-value store built on top of Amazon S3 and known as Dynamo. This paper presents the Dynamo system, which is aware of the data schema it manages. It is a distributed key-value store that is designed to be robust to partial failures (e.g., the loss of a single data center, or the loss of half of the replicas of an item). The design is based on a key value store that is highly available and scalable.

Dynamo uses a synthesis of well known techniques to achieve scalability and availability. Data is partitioned and replicated using consistent hashing [10], and consistency is facilitated by object versioning [12]. The consistency among replicas during updates is maintained by a quorum-like technique and a decentralized replica synchronization protocol. Dynamo employs

[Shapiro et al., SSS '11]

[DeCandia et al., SOSP '07]
Two “styles” of Conflict-Free Replicated Data Types:

“Convergent”
CvRDTs
“state-based”

“Commutative”
CmRDTs
“op-based”

\[
\text{put} \cdot \text{put} = \text{put} \cdot \text{put}
\]
Two “styles” of Conflict-Free Replicated Data Types:

```
CvRDTs
```

```
“Convergent”

CmRDTs
```

```
“Commutative”

“state-based”
```

```
“op-based”
```

[Shapiro et al., SSS ’11]

```
put \cdot put \equiv put \cdot put
```

Two “styles” of Conflict-Free Replicated Data Types:

```
| "Convergent" | "Commumative"
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CvRDTs</td>
<td>CmRDTs</td>
</tr>
<tr>
<td>&quot;state-based&quot;</td>
<td>&quot;op-based&quot;</td>
</tr>
</tbody>
</table>
```

[Shapiro et al., SSS ’11]

```
\[ \text{put} \cdot \text{put} = \text{put} \cdot \text{put} \]
```
## LVars vs. CvRDTs

<table>
<thead>
<tr>
<th>LVars</th>
<th>CvRDTs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold reads</td>
<td>Ordinary reads</td>
</tr>
<tr>
<td>(deterministic)</td>
<td>(nondeterministic)</td>
</tr>
<tr>
<td>Least-upper-bound writes</td>
<td>General inflationary writes</td>
</tr>
<tr>
<td>(every write computes a join)</td>
<td>(only merges must be joins)</td>
</tr>
<tr>
<td>Shared memory</td>
<td>Replicated!</td>
</tr>
</tbody>
</table>
LVars vs. CvRDTs

<table>
<thead>
<tr>
<th>Threshold reads (deterministic)</th>
<th>Ordinary reads (nondeterministic)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least-upper-bound writes</td>
<td>General inflationary writes</td>
</tr>
<tr>
<td>(every write computes a join)</td>
<td>(only merges must be joins)</td>
</tr>
<tr>
<td>Shared memory</td>
<td>Replicated!</td>
</tr>
</tbody>
</table>
## LVars vs. CvRDTs

<table>
<thead>
<tr>
<th>LVars</th>
<th>CvRDTs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Threshold reads</strong>&lt;br&gt;(deterministic)</td>
<td><strong>Ordinary reads</strong>&lt;br&gt;(nondeterministic)</td>
</tr>
<tr>
<td>Least-upper-bound writes&lt;br&gt;(every write computes a join)</td>
<td>General inflationary writes&lt;br&gt;(only <em>merges</em> must be joins)</td>
</tr>
<tr>
<td>Shared memory</td>
<td>Replicated!</td>
</tr>
</tbody>
</table>
LVars vs. CvRDTs

<table>
<thead>
<tr>
<th>Threshold reads (deterministic)</th>
<th>Ordinary reads (nondeterministic)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Least-upper-bound writes (every write computes a join)</td>
<td>General inflationary writes (only merges must be joins)</td>
</tr>
<tr>
<td>Shared memory</td>
<td>Replicated!</td>
</tr>
</tbody>
</table>

so we’re proposing

[WoDet ’14]
Adding threshold reads to CvRDTs

LVars vs. CvRDTs

Threshold reads (deterministic) | Ordinary reads (nondeterministic)
Least-upper-bound writes (every write computes a join) | General inflationary writes (only merges must be joins)
Shared memory | Replicated!

- Adding threshold reads to CvRDTs
- One framework for reasoning about both eventual and strong consistency

so we’re proposing [WoDet ’14]
LVars vs. CvRDTs

- **Threshold reads** (deterministic)
- **Least-upper-bound writes** (every write computes a join)
- **Shared memory**

- **Ordinary reads** (nondeterministic)
- **General inflationary writes** (only merges must be joins)
- **Replicated**

- Adding **threshold** reads to CvRDTs
  - One framework for reasoning about both eventual and strong consistency

- Adding **general inflationary writes** to LVars
  - Non-idempotent, incrementable counters

---

so we're proposing
[WoDet '14]
Parallel systems

Distributed systems
LVars and LVish across the landscape:

- Kahn process networks
- Single-assignment
- Imperative disjoint
- \( \lambda \)-calculus

\[ f \left( g \left( \begin{array}{c} x \\ \hline \end{array} \right) \right) \]

\[ f \left( h \left( \begin{array}{c} y \\ \hline \end{array} \right) \right) \]
LVars and LVish across the landscape:

Kahn process networks

single-assignment

imperative disjoint

\lambda\text{-calculus}

\begin{align*}
  f \left( (g \ x) \ (h \ y) \right)
\end{align*}

CvRDTs

distributed LVish

quasi-det.
LVars and LVish across the landscape:

Kahn process networks

single-assignment

imperative disjoint

λ-calculus

f (g x) (h y)

g(left) h(right)

Thank you!

Email: lkuper@cs.indiana.edu
LVars project repo: github.com/iu-parfunc/lvars
Code from this talk: github.com/lkuper/lvar-examples
Papers: cs.indiana.edu/~lkuper
Research blog: composition.al