Retrofitting Effect Handlers onto OCaml

"KC" Sivaramakrishnan



MADRAS 🐷

Retrofitting Effect Handlers onto OCaml

"KC" Sivaramakrishnan

See PLDI'21 paper



MADRAS 👐

Concurrent Programming

• Computations may be suspended and resumed later

Concurrent Programming

- Computations may be *suspended* and *resumed* later
- Many languages provide concurrent programming mechanisms as primitives
 - ★ async/await JavaScript, Python, Rust, C# 5.0, F#, ...
 - ✦ generators Python, Javascript, …
 - ◆ coroutines C++, Kotlin, Lua, …
 - futures & promises JavaScript, Swift, …

Concurrent Programming

- Computations may be *suspended* and *resumed* later
- Many languages provide concurrent programming mechanisms as primitives
 - ★ async/await JavaScript, Python, Rust, C# 5.0, F#, ...
 - ✦ generators Python, Javascript, …
 - coroutines C++, Kotlin, Lua, …
 - futures & promises JavaScript, Swift, …
- Often include different primitives for concurrent programming
 - JavaScript has async/await, generators, promises, and callbacks!!

Concurrent Programming in OCaml

- No primitive support for concurrent programming in OCaml
 - + Lwt and Async concurrent programming libraries
 - Callback-oriented programming with monadic syntax >>=

Concurrent Programming in OCaml

- No primitive support for concurrent programming in OCaml
 - + Lwt and Async concurrent programming libraries
 - Callback-oriented programming with monadic syntax >>=
- Suffers many pitfalls of callback-oriented programming
 - No backtraces, no exceptions, more closures
 - Monads split the ecosystem into Asynchronous and Synchronous
 - Bob Nystrom, "What colour is your function?"

Effect Handlers

• A mechanism for programming with user-defined effects



- A mechanism for programming with user-defined effects
- Modular and composable basis of non-local control-flow mechanisms
 - Exceptions, generators, lightweight threads, promises, asynchronous IO, coroutines as *libraries*



- A mechanism for programming with user-defined effects
- Modular and composable basis of non-local control-flow mechanisms
 - Exceptions, generators, lightweight threads, promises, asynchronous IO, coroutines as libraries
- Effect handlers ~= first-class, restartable exceptions
 - Structured programming with delimited continuations

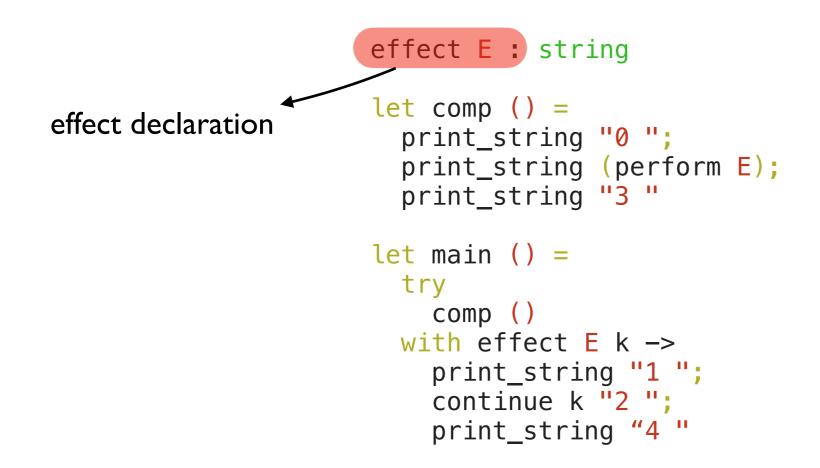
Effect Handlers

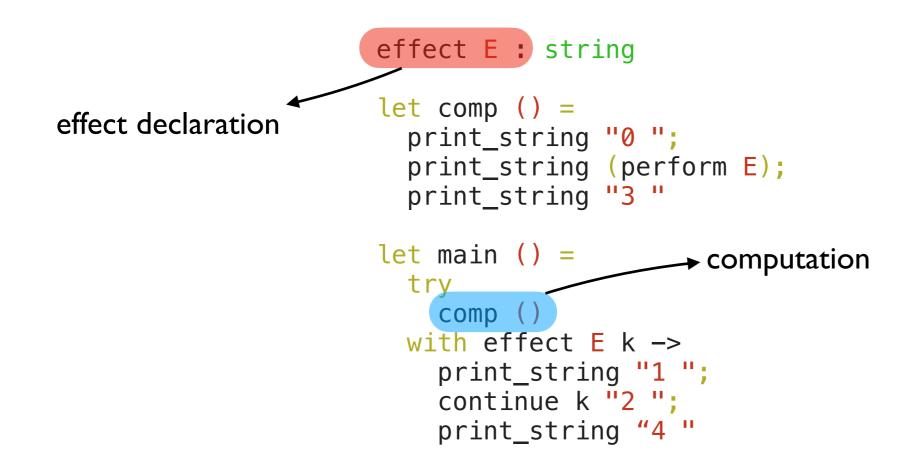
- A mechanism for programming with user-defined effects
- Modular and composable basis of nonmechanisms
 - Exceptions, generators, lightweight thre IO, coroutines as libraries
- Effect handlers ~= first-class, restartab
 - Structured programming with delimited

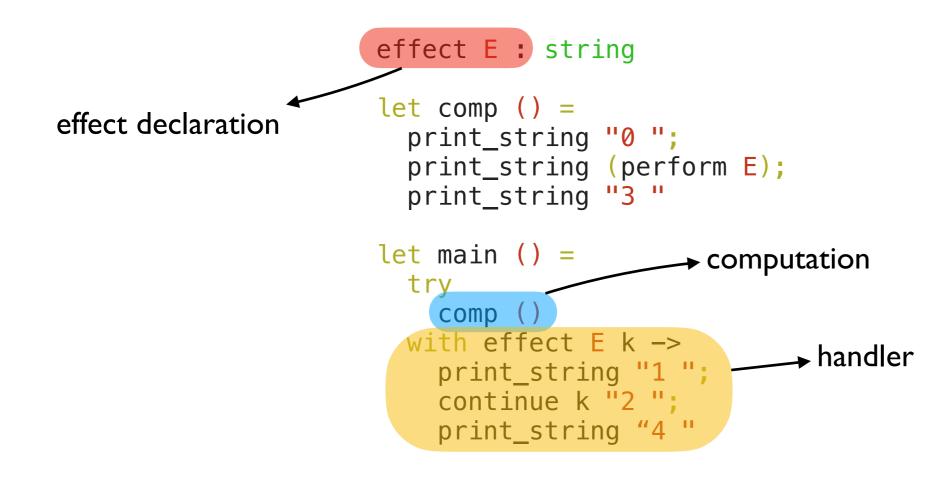
https://github.com/ocaml-multicore/effects-examples

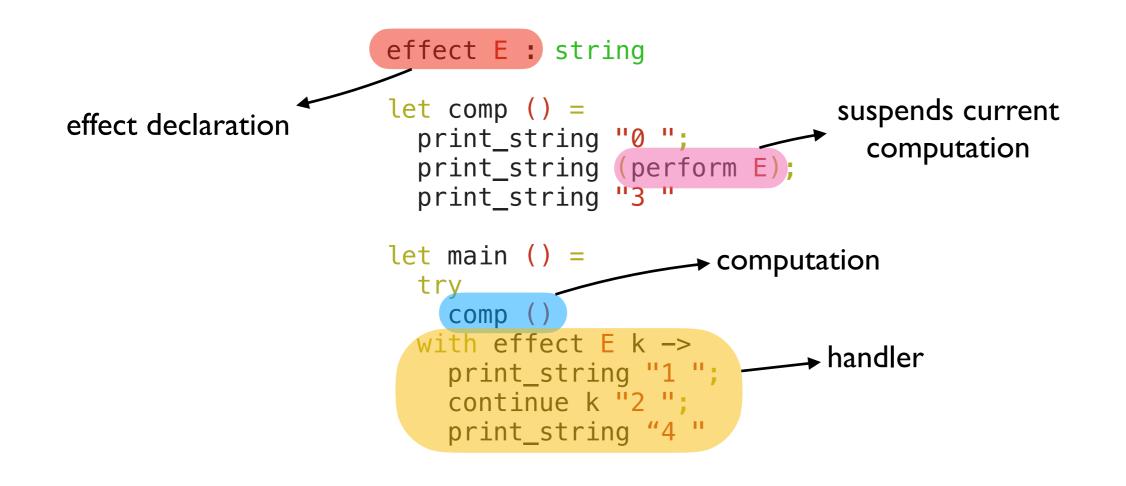
- Direct-style asynchronous I/O
- Generators
- Resumable parsers
- Probabilistic Programming
- Reactive UIs

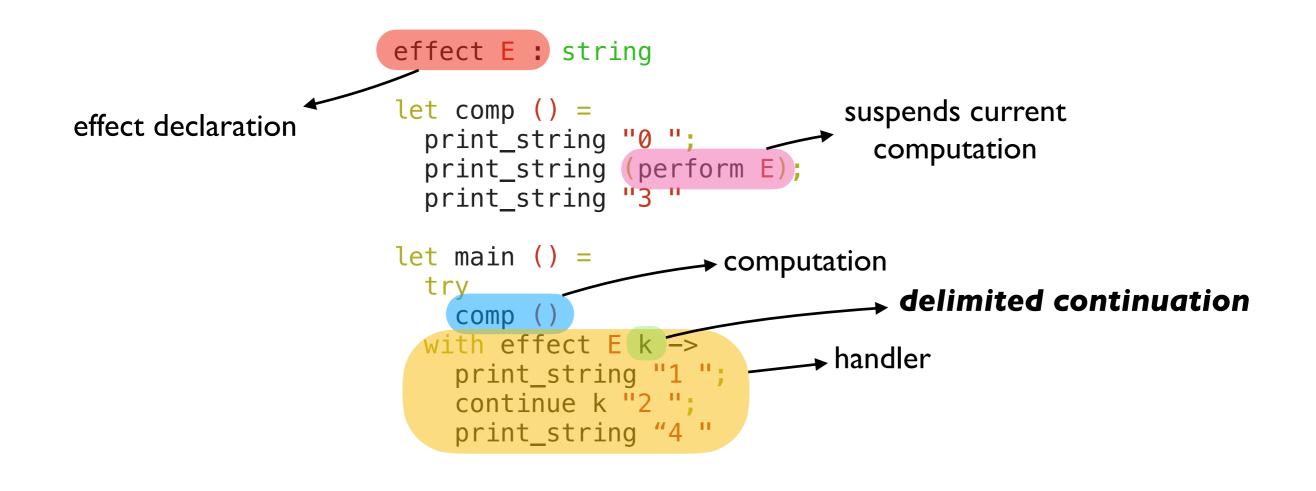
```
effect E : string
let comp () =
    print_string "0 ";
    print_string (perform E);
    print_string "3 "
let main () =
    try
        comp ()
    with effect E k ->
        print_string "1 ";
        continue k "2 ";
        print_string "4 "
```

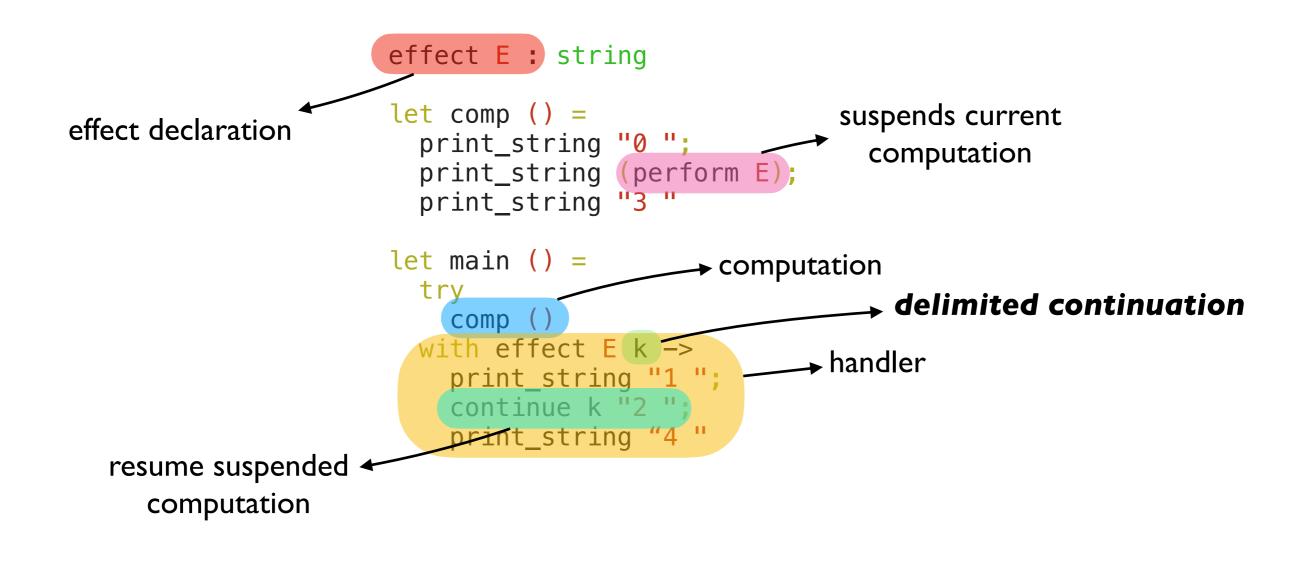


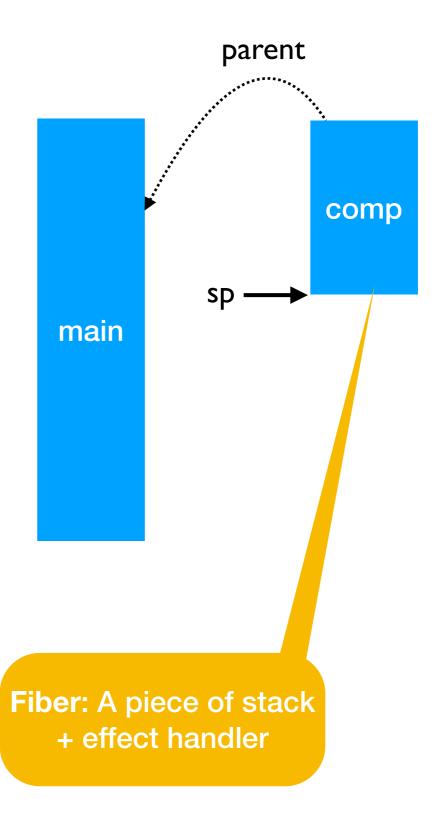




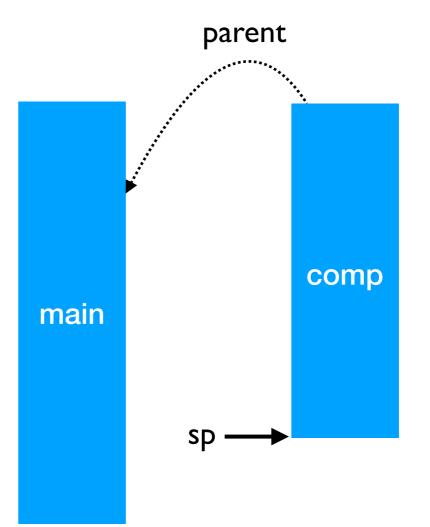




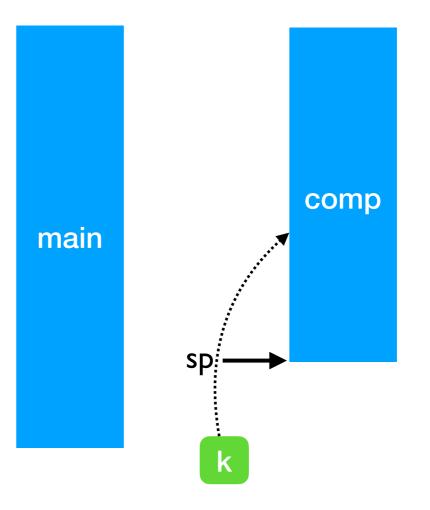




```
effect E : string
let comp () =
    print_string "0 ";
pc    print_string (perform E);
    print_string "3 "
let main () =
    try
    comp ()
with effect E k ->
    print_string "1 ";
    continue k "2 ";
    print_string "4 "
```



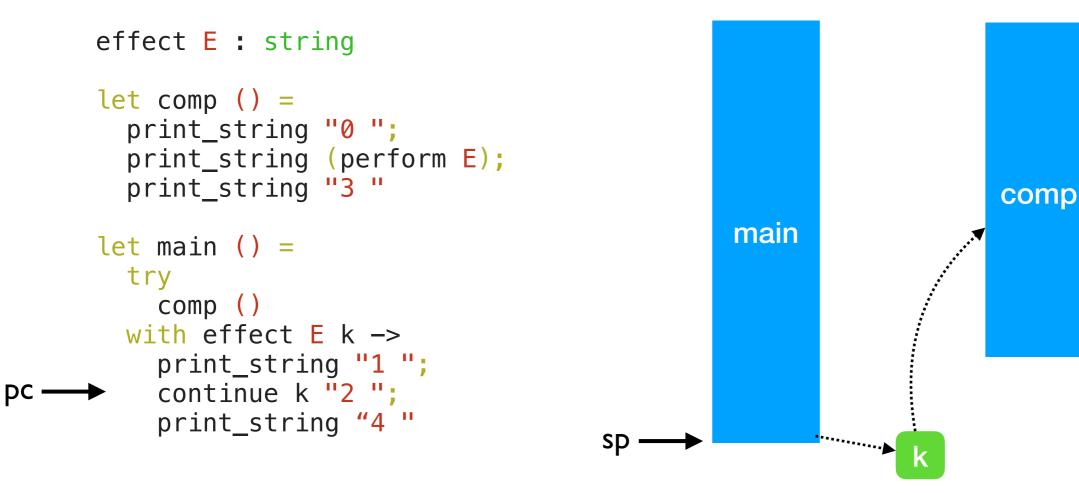
```
effect E : string
let comp () =
    print_string "0 ";
pc >> print_string (perform E);
    print_string "3 "
let main () =
    try
        comp ()
with effect E k ->
        print_string "1 ";
        continue k "2 ";
        print_string "4 "
```



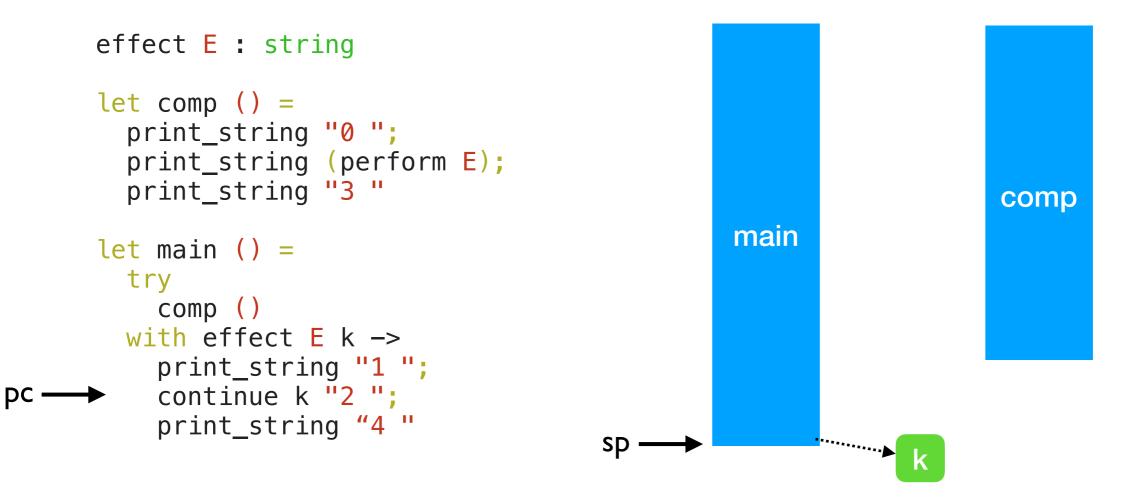
```
effect E : string
      let comp () = 
        print_string "0 ";
    print_string (perform E);
pc -
        print_string "3 "
                                                                   comp
                                                 main
      let main () =
        try
          comp ()
        with effect E k ->
          print_string "1 ";
          continue k "2 ";
          print_string "4 "
                                        sp
```

0

```
effect E : string
      let comp () = 
        print_string "0 ";
        print_string (perform E);
        print_string "3 "
                                                                    comp
                                                  main
      let main () =
        try
          comp ()
        with effect E k ->
          print_string "1 ";
pc –
          continue k "2 ";
          print_string "4 "
                                         sp
```



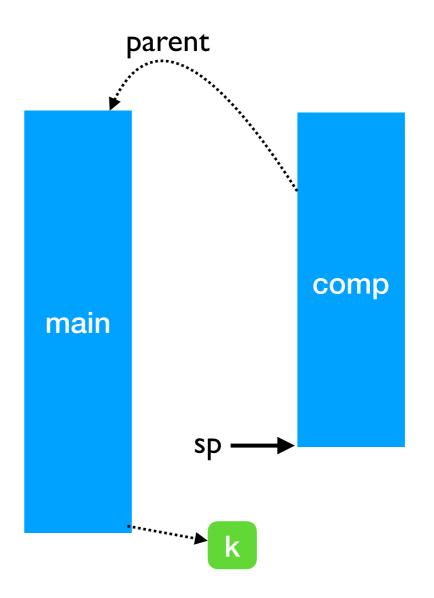
0 |



0 |

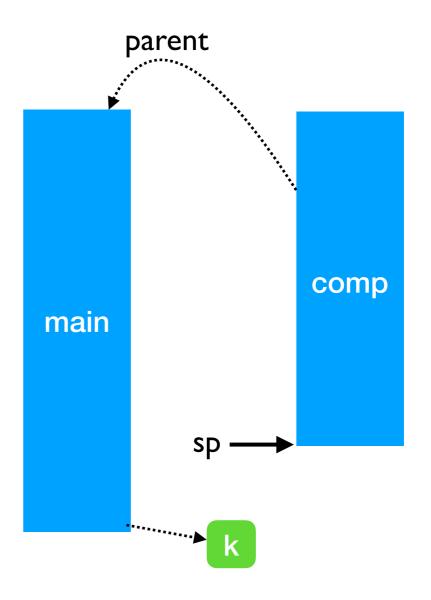
```
effect E : string
let comp () =
    print_string "0 ";
    print_string (perform E);
    print_string "3 "
let main () =
    try
        comp ()
with effect E k ->
        print_string "1 ";
        continue k "2 ";
        print_string "4 "
```

pc -



0 1

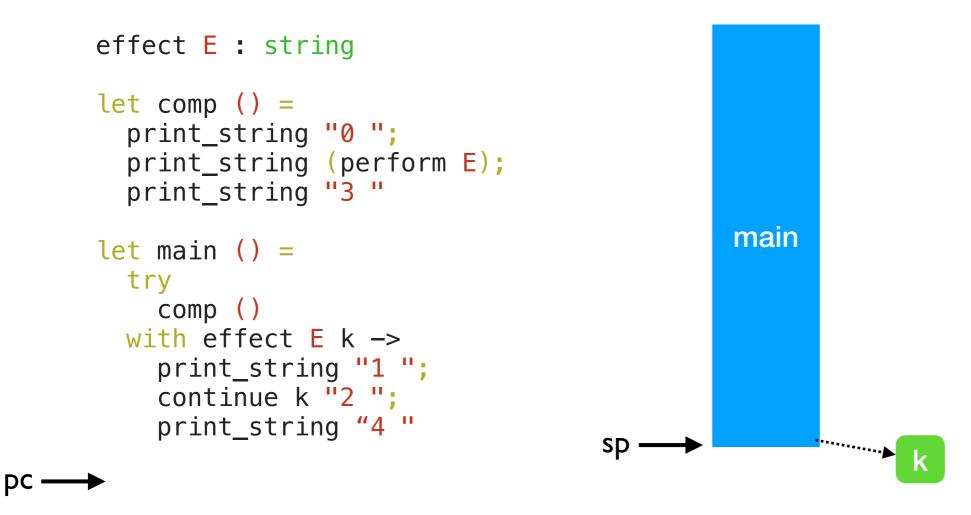
```
effect E : string
let comp () =
    print_string "0 ";
    print_string (perform E);
pc 		 print_string "3 "
let main () =
    try
        comp ()
with effect E k ->
        print_string "1 ";
        continue k "2 ";
        print_string "4 "
```



0 I 2

```
effect E : string
       let comp () = 
         print_string "0 ";
         print_string (perform E);
         print_string "3 "
                                                       main
       let main () =
         try
           comp ()
         with effect E k \rightarrow
           print_string "1 ";
           continue k "2 ";
           print_string "4"
рC
                                                              ••••••••••••••••
                                             SD .
```

0 I 2 3



0 I 2 3 4

Lightweight Threading

effect Fork : (unit -> unit) -> unit
effect Yield : unit

Lightweight Threading

```
effect Fork : (unit -> unit) -> unit
effect Yield : unit
let run main =
  ... (* assume queue of continuations *)
  let run_next () =
   match dequeue () with
    Some k -> continue k ()
     None -> ()
  in
  let rec spawn f =
    match f () with
     () -> run_next () (* value case *)
    | effect Yield k -> enqueue k; run_next ()
    effect (Fork f) k -> enqueue k; spawn f
  in
  spawn main
```

Lightweight Threading

```
effect Fork : (unit -> unit) -> unit
effect Yield : unit
let run main =
  ... (* assume queue of continuations *)
  let run_next () =
   match dequeue () with
    Some k -> continue k ()
     None -> ()
  in
  let rec spawn f =
   match f () with
     () -> run_next () (* value case *)
    effect Yield k -> enqueue k; run_next ()
    effect (Fork f) k -> enqueue k; spawn f
  in
  spawn main
let fork f = perform (Fork f)
```

```
let yield () = perform Yield
```

Lightweight threading

```
let main () =
  fork (fun _ -> print_endline "1.a"; yield (); print_endline "1.b");
  fork (fun _ -> print_endline "2.a"; yield (); print_endline "2.b")
;;
run main
```

Lightweight threading

```
let main () =
  fork (fun _ -> print_endline "1.a"; yield (); print_endline "1.b");
  fork (fun _ -> print_endline "2.a"; yield (); print_endline "2.b")
;;
run main
```

```
1.a
2.a
1.b
2.b
```

Lightweight threading

```
let main () =
  fork (fun _ -> print_endline "1.a"; yield (); print_endline "1.b");
  fork (fun _ -> print_endline "2.a"; yield (); print_endline "2.b")
;;
run main
```

- 1.a 2.a 1.b 2.b
- Direct-style (no monads)
- User-code need not be aware of effects
- No Async vs Sync distinction

Retrofitting Challenges

- Millions of lines of legacy code
 - Written without non-local control-flow in mind
 - Cost of refactoring sequential code itself is prohibitive

Retrofitting Challenges

- Millions of lines of legacy code
 - Written without non-local control-flow in mind
 - Cost of refactoring sequential code itself is prohibitive
- OCaml uses the same system stack for both OCaml and C
 - Fast exceptions and FFI between C and OCaml
 - No stack overflow checks needed
 - Excellent compatibility with debugging (gdb) and profiling (perf) tools

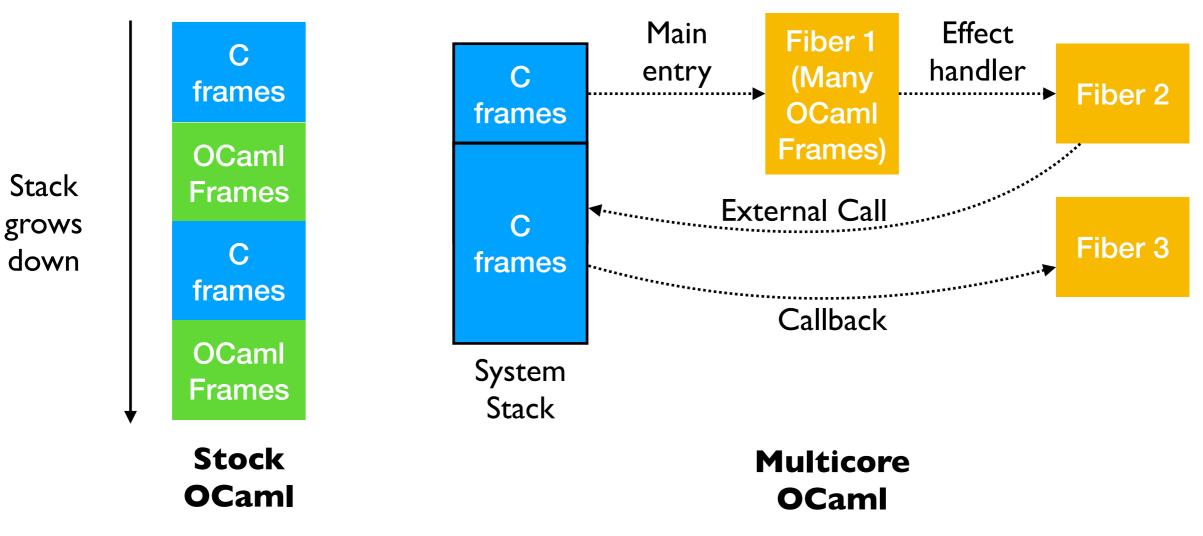
Retrofitting Challenges

- Millions of lines of legacy code
 - Written without non-local control-flow in mind
 - Cost of refactoring sequential code itself is prohibitive
- OCaml uses the same system stack for both OCaml and C
 - Fast exceptions and FFI between C and OCaml
 - No stack overflow checks needed
 - + Excellent compatibility with debugging (gdb) and profiling (perf) tools

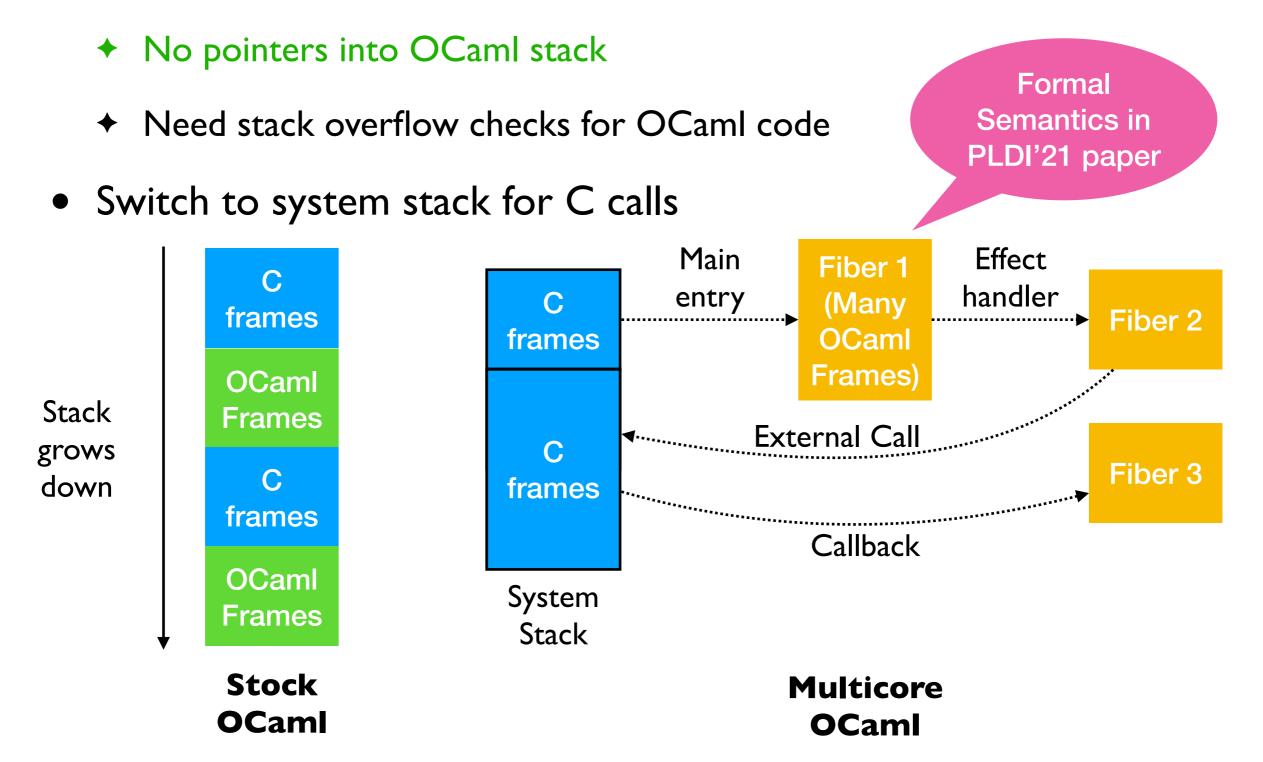
Must preserve feature, tooling, performance compatibility

- A stack of runtime-managed, dynamically growing stack segments
 - No pointers into OCaml stack
 - Need stack overflow checks for OCaml code

- A stack of runtime-managed, *dynamically growing* stack segments
 - No pointers into OCaml stack
 - Need stack overflow checks for OCaml code
- Switch to system stack for C calls



• A stack of runtime-managed, *dynamically growing* stack segments



Switching stacks fast

• One-shot — capture and resumption does not involve copying frames

Switching stacks fast

- One-shot capture and resumption does not involve copying frames
- No callee-saved registers in OCaml
 - Switching between stacks need not save & restore register state

```
let foo () =
   (* a *)
   try
    (* b *)
    perform E
    (* d *)
   with effect E k ->
    (* c *)
    continue k ()
    (* e *)
```

```
let foo () =
   (* a *)
   try
    (* b *)
    perform E
    (* d *)
with effect E k ->
    (* c *)
    continue k ()
    (* e *)
```

Instruction Sequence	Significance	
a to b	Create a new stack & run the computation	
b to c	Performing & handling an effect	
c to d	Resuming a continuation	
d to e	Returning from a computation & free the stack	

• Each of the instruction sequences involves a stack switch

```
let foo () =
   (* a *)
   try
    (* b *)
    perform E
    (* d *)
   with effect E k ->
    (* c *)
    continue k ()
    (* e *)
```

Instruction Sequence	Significance	
a to b	Create a new stack & run the computation	
b to c	Performing & handling an effect	
c to d	Resuming a continuation	
d to e	Returning from a computation & free the stack	

- Each of the instruction sequences involves a stack switch
- Intel(R) Xeon(R) Gold 5120 CPU @ 2.20GHz
 - Cost measured using Intel PT's cycle accurate tracing
 - For calibration, memory read latency is **90 ns** (local NUMA node) and **I45 ns** (remote NUMA node)

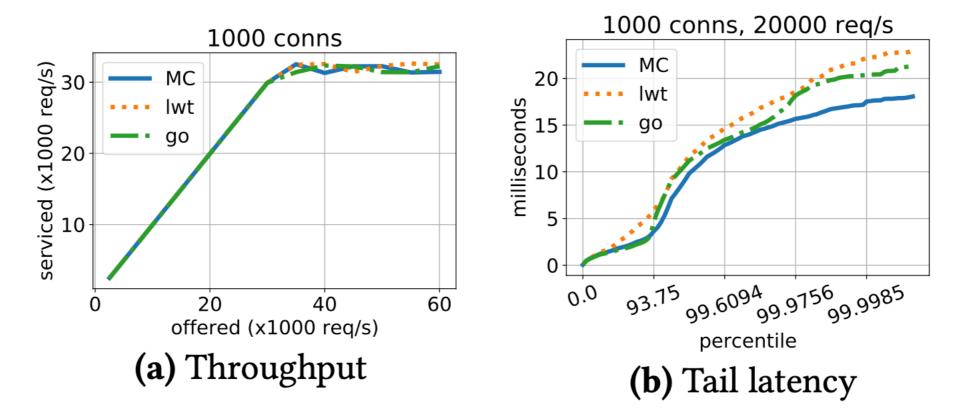
```
let foo () =
   (* a *)
   try
    (* b *)
    perform E
    (* d *)
with effect E k ->
    (* c *)
    continue k ()
    (* e *)
```

Instruction Sequence	Significance	Time (ns)
a to b	Create a new stack & run the computation	23
b to c	Performing & handling an effect	5
c to d	Resuming a continuation	11
d to e	Returning from a computation & free the stack	7

- Each of the instruction sequences involves a stack switch
- Intel(R) Xeon(R) Gold 5120 CPU @ 2.20GHz
 - Cost measured using Intel PT's cycle accurate tracing
 - For calibration, memory read latency is 90 ns (local NUMA node) and I45 ns (remote NUMA node)

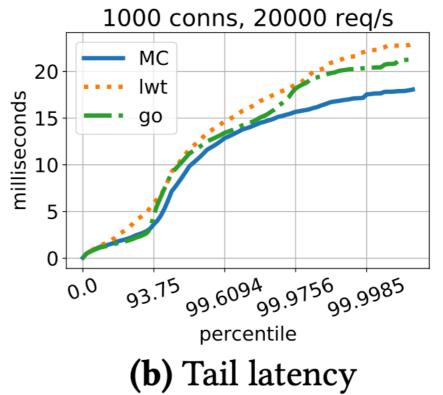
- Effect handlers for asynchronous I/O in direct-style
 - https://github.com/kayceesrk/ocaml-aeio
- Variants
 - **Go** + net/http (GOMAXPROCS=I)
 - OCaml + http/af + Lwt (explicit callbacks)
 - OCaml + http/af + Effect handlers (MC)
- Performance measured using wrk2

- Effect handlers for asynchronous I/O in direct-style
 - https://github.com/kayceesrk/ocaml-aeio
- Variants
 - **Go** + net/http (GOMAXPROCS=I)
 - OCaml + http/af + Lwt (explicit callbacks)
 - OCaml + http/af + Effect handlers (MC)
- Performance measured using wrk2

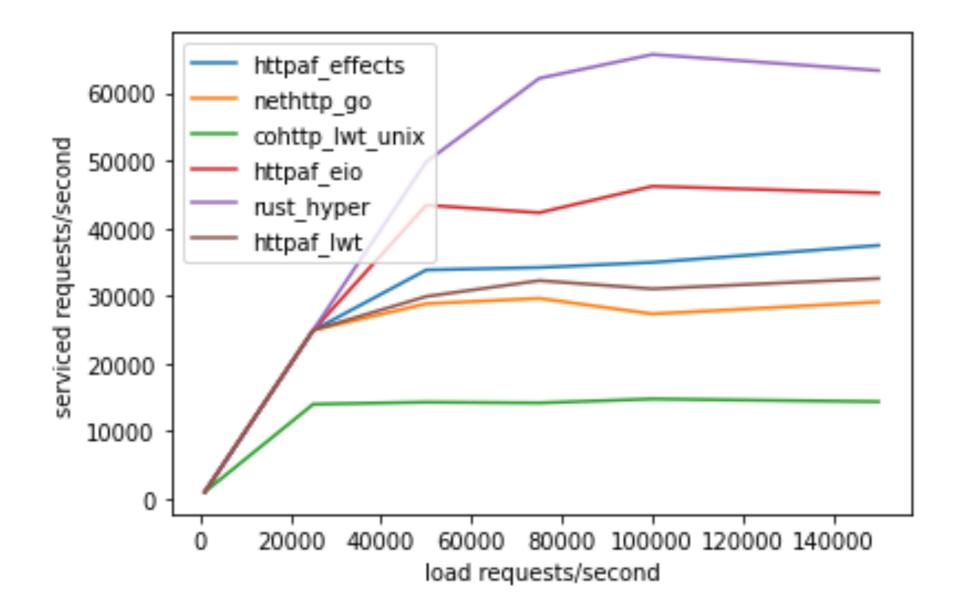


- Effect handlers for asynchronous I/O in direct-style
 - https://github.com/kayceesrk/ocaml-aeio
- Variants
 - Go + net/http (GOMAXPROCS=I)
 - OCaml + http/af + Lwt (explicit callbacks)
 - OCaml + http/af + Effect handlers (MC)
- Performance measured using wrk2
 - 1000 conns serviced (x1000 req/s) MC 20 30 milliseconds lwt go 20 5 10 60 20 0 40 offered (x1000 reg/s) (a) Throughput

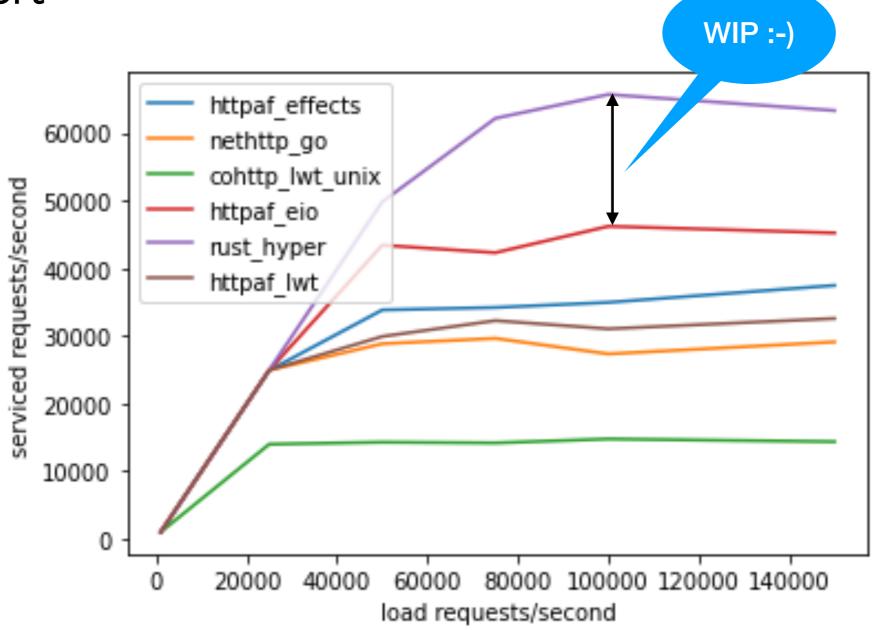
- Direct style (no monadic syntax)
- Can use OCaml exceptions!
- Backtrace per thread (request)
- gdb & perf work!



eio: effects-based I/O over Linux kernel's new io_uring support



eio: effects-based I/O over Linux kernel's new io_uring support



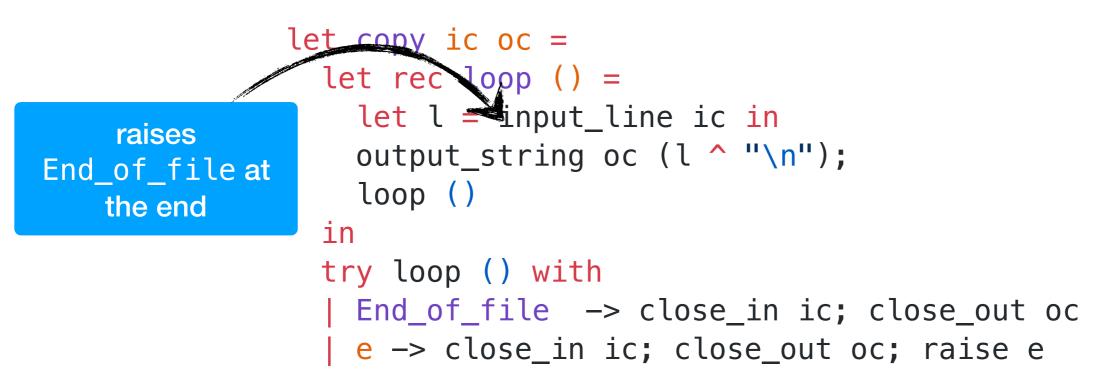
- OCaml is a systems programming language
 - Manipulates resources such as files, sockets, buffers, etc.

- OCaml is a systems programming language
 - Manipulates resources such as files, sockets, buffers, etc.
- OCaml code is written in <u>defensive style</u> to guard against exceptional behaviour and clear up resources

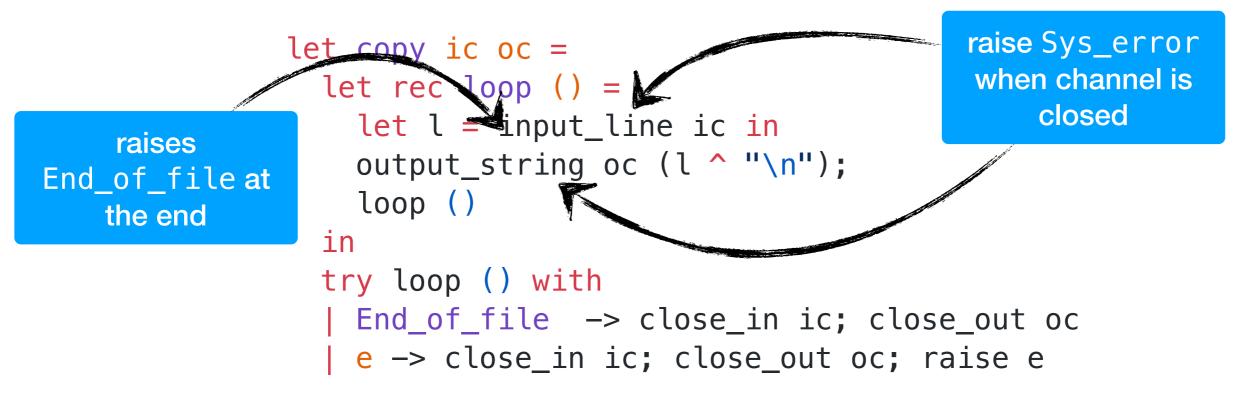
- OCaml is a systems programming language
 - Manipulates resources such as files, sockets, buffers, etc.
- OCaml code is written in *defensive style* to guard against exceptional behaviour and clear up resources

```
let copy ic oc =
   let rec loop () =
      let l = input_line ic in
      output_string oc (l ^ "\n");
      loop ()
   in
   try loop () with
      [ End_of_file -> close_in ic; close_out oc
      [ e -> close_in ic; close_out oc; raise e
```

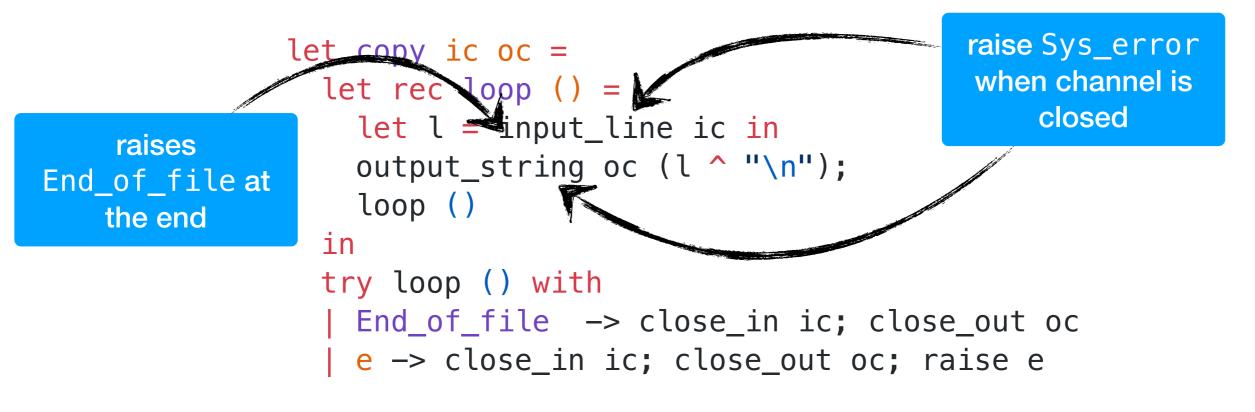
- OCaml is a systems programming language
 - Manipulates resources such as files, sockets, buffers, etc.
- OCaml code is written in *defensive style* to guard against exceptional behaviour and clear up resources



- OCaml is a systems programming language
 - Manipulates resources such as files, sockets, buffers, etc.
- OCaml code is written in *defensive style* to guard against exceptional behaviour and clear up resources



- OCaml is a systems programming language
 - Manipulates resources such as files, sockets, buffers, etc.
- OCaml code is written in *defensive style* to guard against exceptional behaviour and clear up resources



We would like to make this code transparently asynchronous

effect In_line : in_channel -> string
effect Out_str : out_channel * string -> unit

effect In_line : in_channel -> string
effect Out_str : out_channel * string -> unit

let input_line ic = perform (In_line ic)
let output_string oc s = perform (Out_str (oc,s))

```
effect In_line : in_channel -> string
effect Out_str : out_channel * string -> unit
let input_line ic = perform (In_line ic)
let output_string oc s = perform (Out_str (oc,s))
let run aio f = match f () with
∨ -> ∨
| effect (In_line chan) k ->
    register_async_input_line chan k;
    run_next ()
effect (Out_str (chan, s)) k ->
    register_async_output_string chan s k;
    run next ()
```

```
effect In_line : in_channel -> string
effect Out_str : out_channel * string -> unit
let input_line ic = perform (In_line ic)
let output_string oc s = perform (Out_str (oc,s))
let run aio f = match f () with
 V -> V
 effect (In_line chan) k ->
    register_async_input_line chan k;
    run_next ()
effect (Out_str (chan, s)) k ->
    register_async_output_string chan s k;
    run next ()
```

Continue with appropriate value when the asynchronous IO call returns

```
effect In_line : in_channel -> string
effect Out_str : out_channel * string -> unit
let input_line ic = perform (In_line ic)
let output_string oc s = perform (Out_str (oc,s))
let run_aio f = match f () with
 V -> V
 effect (In_line chan) k ->
    register_async_input_line chan k;
    run_next ()
effect (Out_str (chan, s)) k ->
    register_async_output_string chan s k;
    run next ()
```

- Continue with appropriate value when the asynchronous IO call returns
- But what about termination? End_of_file and Sys_error exceptional cases.

Discontinue

discontinue k End_of_file

- We add a discontinue primitive to resume a continuation by raising an exception
- On End_of_file and Sys_error, the asynchronous IO scheduler uses discontinue to raise the appropriate exception

- Resources such as sockets, file descriptors, channels and buffers are *linear* resources
 - Created and destroyed exactly once

- Resources such as sockets, file descriptors, channels and buffers are *linear* resources
 - Created and destroyed exactly once
- OCaml functions return exactly once with value or exception
 - Defensive programming already guards against exceptional return cases

- Resources such as sockets, file descriptors, channels and buffers are *linear* resources
 - Created and destroyed exactly once
- OCaml functions return exactly once with value or exception
 - Defensive programming already guards against exceptional return cases
- With effect handlers, functions may return *at-most once* if continuation not resumed
 - This breaks resource-safe legacy code

effect E : unit
let foo () = perform E

```
effect E : unit
let foo () = perform E
let bar () =
  let ic = open_in "input.txt" in
  match foo () with
  | v -> close_in ic
  | exception e -> close_in ic; raise e
```

Linearity

```
effect E : unit
let foo () = perform E
let bar () =
  let ic = open_in "input.txt" in
  match foo () with
  | v -> close_in ic
  | exception e -> close_in ic; raise e
let baz () =
  try bar () with
  | effect E _ -> () (* leaks ic *)
```

Linearity

```
effect E : unit
let foo () = perform E
let bar () =
  let ic = open_in "input.txt" in
  match foo () with
  | v -> close_in ic
  | exception e -> close_in ic; raise e
let baz () =
  try bar () with
  | effect E _ -> () (* leaks ic *)
```

We assume that captured continuations are resumed exactly once either using continue or discontinue

- OCaml has excellent compatibility with debugging and profiling tools gdb, lldb, perf, libunwind, etc.
 - DWARF stack unwinding support

- OCaml has excellent compatibility with debugging and profiling tools gdb, lldb, perf, libunwind, etc.
 - DWARF stack unwinding support
- Multicore OCaml supports DWARF stack unwinding across fibers

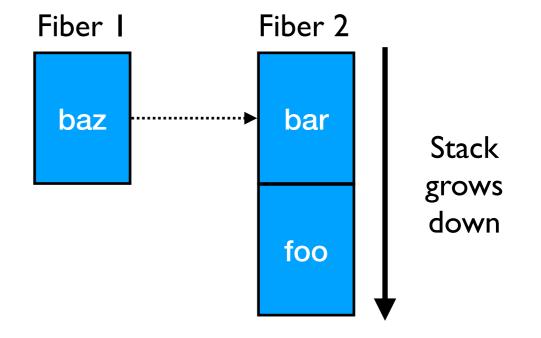
- OCaml has excellent compatibility with debugging and profiling tools gdb, lldb, perf, libunwind, etc.
 - DWARF stack unwinding support
- Multicore OCaml supports DWARF stack unwinding across fibers

```
effect E : unit
let foo () = perform E
let bar () =
  let ic = open_in "input.txt" in
  match foo () with
  | v -> close_in ic
  | exception e ->
      close_in ic; raise e
let baz () =
  try bar () with
  | effect E _ -> () (* leak *)
```

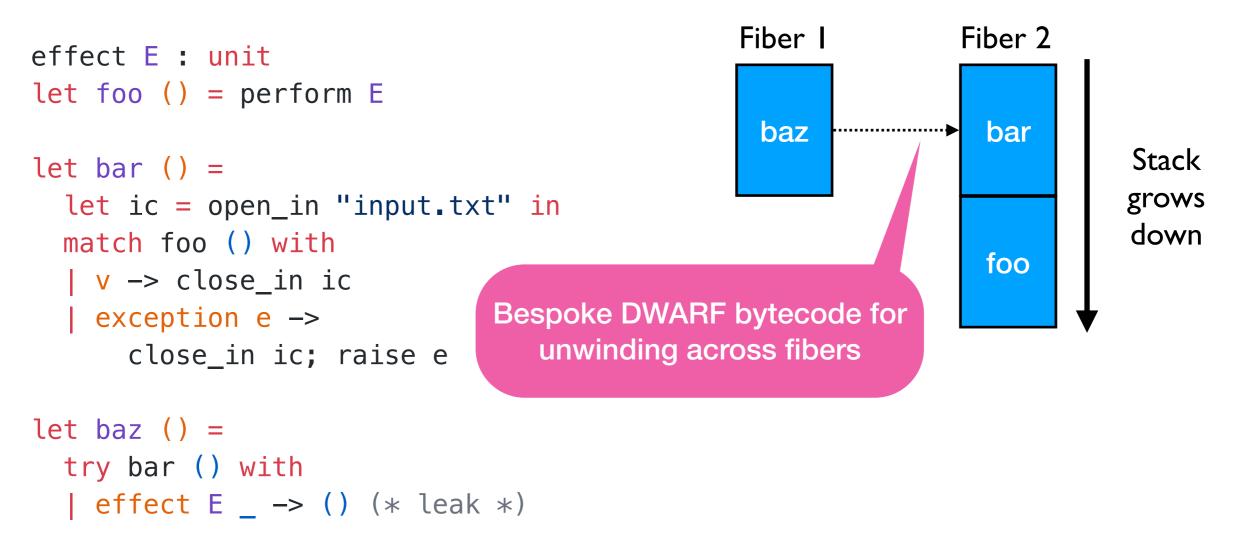
- OCaml has excellent compatibility with debugging and profiling tools gdb, lldb, perf, libunwind, etc.
 - DWARF stack unwinding support
- Multicore OCaml supports DWARF stack unwinding across fibers

```
effect E : unit
let foo () = perform E
let bar () =
  let ic = open_in "input.txt" in
  match foo () with
  | v -> close_in ic
  | exception e ->
      close_in ic; raise e
let baz () =
  try bar () with
```

| effect E _ -> () (* leak *)



- OCaml has excellent compatibility with debugging and profiling tools gdb, lldb, perf, libunwind, etc.
 - DWARF stack unwinding support
- Multicore OCaml supports DWARF stack unwinding across fibers



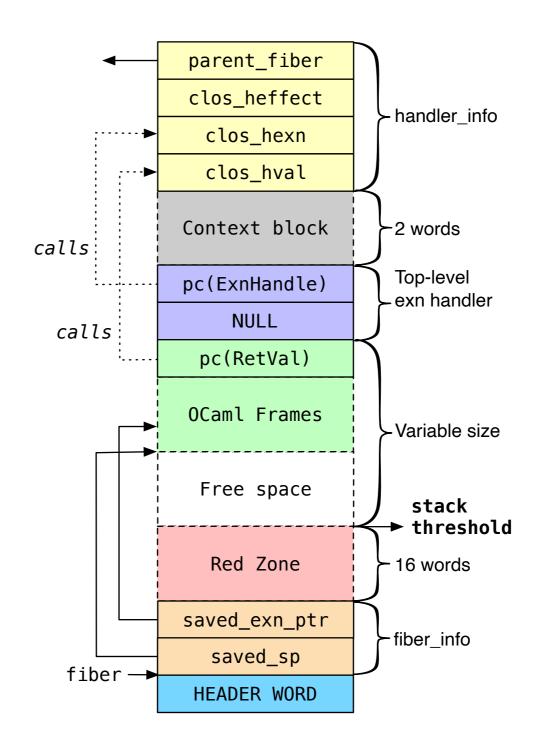
```
(lldb) bt
effect E : unit
let foo () = perform E
                                  * thread #1, name = 'a.out', stop reason = ...
                                    * #0: 0x58b208 caml_perform
let bar () =
                                      #1: 0x56aa5d camlTest__foo_83 at test.ml:4
  let ic = open_in "input.txt" in
                                      #2: 0x56aae2 camlTest__bar_85 at test.ml:9
  match foo () with
                                      #3: 0x56a9fc camlTest__fun_199 at test.ml:14
  ∨ -> close in ic
                                      #4: 0x58b322 caml_runstack + 70
  | exception e ->
                                      #5: 0x56ab99 camlTest__baz_91 at test.ml:14
      close_in ic; raise e
                                      #6: 0x56ace6 camlTest__entry at test.ml:21
                                      #7: 0x56a41c caml_program + 60
let baz () =
 try bar () with
                                      #8: 0x58b0b7 caml_start_program + 135
  effect E _ -> () (* leak *)
                                      #9: ...
```

Thanks!

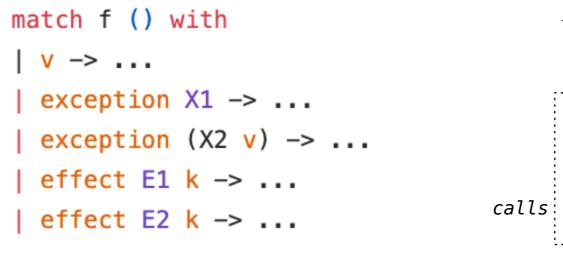
- Multicore OCaml
 - https://github.com/ocaml-multicore/ocaml-multicore
- Effects Examples
 - https://github.com/ocaml-multicore/effects-examples
- Sivaramakrishnan et al, "<u>Retrofitting Effect Handlers onto OCaml</u>", PLDI 2021
 - https://arxiv.org/abs/2104.00250

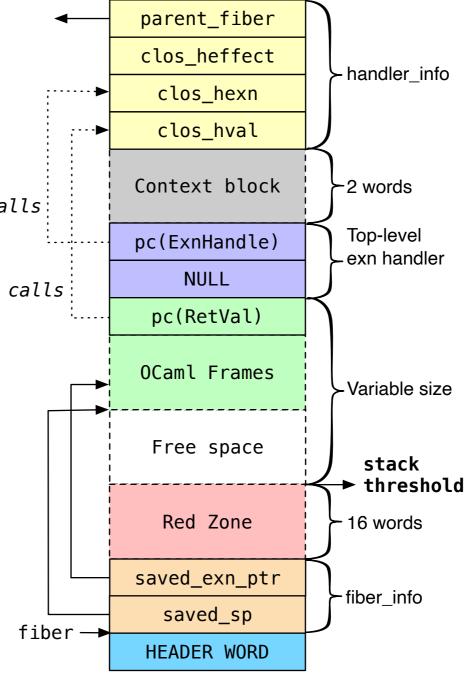
Bonus Slides

Fiber Layout



Fiber Layout





"KC" Sivaramakrishnan

- Who am I?
 - + Asst Prof at IIT Madras, India
 - + Lead the development of Multicore OCaml project
- Interested in learning
 - Compiling effect handlers for uncooperative environments (Wasm, Java, C, JavaScript)
 - Pragmatic effect systems
 - New use cases for effects
- Talks
 - Retrofitting effect handlers onto OCaml (30 minutes)
 - ParaFuzz: Fuzzing Multicore OCaml programs (15 minutes)